Format of Report

* Write-up of company background
  + Take from proposal
* Problem(s)/Issue(s)
  + Take from proposal
  + Expectations from Client
* Project Timeline & implementation (how & why we did things)
  + Process & Approach (Jerome & Gavin)
  + Cleaning & analysing data(Jerome & Gavin)
  + Model building (Jerome & Gavin)
  + Development of visualisations (Ken)
* Model Development (Jerome & Gavin)
  + Deciding factors & considerations
  + Model Explanation & Results obtained - All models obtained
    - Results for all 6 models & Analysis based on train data
    - Results obtained from Test data & Analysis
    - Why does the output variables vary between models
  + Comparison of model results
    - Final model obtained
  + Limitation of Model & Suggestion for improvement
    - Alternative Model Generation
    - Comparison of all new models
  + Justification on why your design considerations are consistent with the problems/issues identified (if any)
    - Rationale for model choice
    - Why we rejected other models
    - Why the model works in meeting objectives
* Visualization of Output (Ken & Balkis)
  + Choice of Software
    - Considerations - Price, Accessibility, Functionality,
    - Recommendations
  + How we organized data on visualizations
    - For each visualization
  + Screenshots - Walkthrough of your demonstration using screen captures (if any)
* Recommendation for STA on Implementation & handover
  + Improvements to data collection (how it can be done better, what else to collect)
  + Schema refinement for appropriate analysis
  + Targeted approach towards reducing overheads
  + Handover procedures (platform, documents)
* Bibliography Citations of research papers, books and periodicals referenced
* Appendix
  + Screenshots of Final model - Walkthrough of your demonstration using screen captures (if any) - Appendix

|  |  |  |
| --- | --- | --- |
| **bg.a** | **Initial Project Objective** | **Synergy between shop floor features to provide BEST-ESITMATE of LOH/pair** |
| **bg.b** | **Eventual success target indicator to be improved upon implementation on the shop floor** | **Provide link between Operations & Financial**  **Understand how Shop Floor parameters affects the Financial indicators (e.g. PPH of 1.08, RFT of 94.6% = what to my Labour, Overheads, and Direct Materials costs per pair)** |
| **1** | **Data Cleansing** | * **Assumptions made for aberrant data (e.g. missing, illogical, outliers etc.)** * **Final cleansed data set: how data was cleaned and rearranged to feed into analysis** |
| **2** | **Data Analysis Method** | * **Which exact months of data used as training** * **What statistical calculation made to select features to go into analysis, what was the cut-off assumption** * **Which models, evaluate reasons for selecting the model** * **How results from each model can be interpreted, how results were reconciled (Can be verbally mentioned)** |
| **3** | **Data Insights** | * **Final suggested algorithm, & boundaries for algo to work** * **How results can be improved e.g. Recommending the proposed data schema** * **How machine learning can be applied for continuous optimisation automatically\*\*\*** |
| **4** | **Visualisation** | * **Comparison between a few visualisation tools (e.g. a simple matrix on functions, learning curve, costs, on cloud flexibility etc.)** * **Evaluation and final recommended visualisation tool** * **Sample of what type of data can be displayed (already in ur ppt)** |

**Executive Summary**

Scope and Objective of the Project

Star Asia Trading (SAT) is an urban shoe manufacturer headquartered in Singapore, with operations primarily based in Indonesia and China. SAT has employed the assistance of the group to employ data analytics aimed to effectively utilise shop-floor data that has been collected from their factories. Specifically, the company requires a machine learning (ML) model that can accurately predict overhead costs and identify the key parameters that are significant to the prediction. Additionally, SAT is looking to create a visualisation tool that provides an effective visual representation of shop-floor parameters.

This report provides a comprehensive discussion on the group’s approach towards SAT’s problem statement the thought processes that drive our project approach. Throughout the report, we share key findings that are relevant to SAT and the insights from the ML models and visualisation tools that can facilitate management decisions. The project approach can be summarized in four main phases - data pre-processing, development of ML models, development of visualizations and further recommendations.

Data Pre-processing

This phase entailed data cleaning and exploratory data analysis. We detail the key processes behind the data cleaning stage, how we structured the data schema and explain the business logic used to select variables for the ML model development phase. Additionally, exploratory data analysis was conducted to achieve a preliminary understanding of the data, to observe the relationships between selected variables and overheads costs.

Model Development

Three algorithms were explored in the model development phase - Linear Regression, Lasso and Random Forest. The group presents an evaluation of the trade-offs between the three models and found Random Forest to be the most optimal algorithm for SAT’s purposes, due to its ability produce accurate predictions and to handle datasets with high dimensionalities. Using the Random Forest model, we explore various approaches (e.g. predicting monthly or daily overheads) taken to determine the most effective model in accurately predicting overheads and identifying key shop-floor parameters.

However, the initial approach had several limitations, which restricted the usefulness of the resulting models. As such, the group further refined the models by restructuring the data schema into non-aggregated variables and focused on predicting variable overheads instead of total overheads. The models that followed were significantly more accurate and effective in identifying variables that were more logically related to predicting overheads.

We propose two final models to SAT. The first is primarily used to predict variable overheads and the second allows management to identify specific areas (e.g. Plant or Cell) in which shop-floor parameters can be optimised.

Visualization Development

The visualizations were created with two primary objectives in mind - to create a useful, consolidated visual representation of SAT’s data and to allow for easy identification of trends and outliers.

This report provides a comparison of various visualization softwares that are available in the market (Tableau, Microstrategy, Microsoft Power BI and Qlik Sense). Based on an evaluation matrix that weighs software usability, pricing and ease-of-use, we find that Tableau is the most appropriate software for SAT’s needs. Finally, we provide a mock-up of different visualisations that SAT can create using Tableau, with a focus on the insights that these dashboards will provide to facilitate decision-making.

Further Recommendations

Finally, we provide further recommendations that SAT can consider. Going forward, artificial neural networks can be a possible ML model to predict overhead costs, due to its suitability of use in manufacturing processes. We also suggest to SAT the use of RFID tags in their production lines, to improve the data collection processes within the factory.

> talk about problem statement

> approach towards data cleanings & exploratory data analysis

> model development

> monthly & daily

> types of models

> best model

> refinements for model

> non-aggregated analysis

> variable OH

>Visualisations

> visualisation software comparison

> dashboard + objectives

> going forward

> ANN

> RFID Tags

> limitations

# 1.0 Company Background

The company engaged in this project is Star Asia Trading (SAT), an urban shoe manufacturer and a wholesale distributor of men’s, women’s and children’s footwear, with major clients like Adidas, Nike and Zara. Currently headquartered in Singapore, with its manufacturing plants located in China and Indonesia, the company aims to modernize the role of a manufacturer in the industry. With a business mode, centred around innovation and data analytics, the company aims to establish itself as the first urban smart shoe factory with a vision to bring manufacturers, consumers and designers together. In the pursuit of these objectives, the company has approached us to assist them in digitally enhancing their existing processes, specifically focusing on the operational efficiency on their shop floor.

In its manufacturing process, SAT’s factory encounters numerous cost variables that are subject to frequent fluctuations, which impacts its operational costs. These cost variables (e.g. percent accuracy of material feeding, pairs per person per hour) are tracked and monitored individually by the factory manager. In general, the company expects that shop floor parameters (e.g. number of downtime periods and man-hours spent) will correspondingly reflect the business’s eventual financial performance. However, the correlation is ambiguous, with some variables contributing more significantly in impacting manufacturing costs than others, while some variables are not significant at all.

At the same time, the company experiences information scarcity as not all variables & quantitative inputs from the shop floor are consistently available to establish a fixed model equation, to determine costs. While Star Asia Trading has been collecting considerable amounts of data regarding its shop floor operations, it is currently under-utilized as there are no formal processes in place to analyse the data collected.

Through their engagement with the team, the company aims to enhance their ability to make decisions with the employment of data analytics tools. SAT has provided the team with historical data of shop floor parameters and financial data, in order to predict its Cost Breakdown (CBD) for the company’s shop floor, based on known variable inputs.

# 2.0 Project Statement

Based on the discussions with SAT, the company is seeking to identify a model where a combination of known variables observed on the shop floor are utilized to accurately predict overhead costs. As a result, the company strives to obtain a more complete picture of their operations, by effective utilization of the data collected from the shop floor. They have requested for our group to develop this model through an open source platform as well as a visualisation tool which provides a useful representation of shop floor parameters.

In the context of this study, the focus of the project will be to estimate overhead costs per pair of shoes manufactured and determine the specific variables that can be used to make this estimation as accurately as possible; with a minimum of 60% accuracy.

## 2.1 Project Deliverables

Below are the breakdown of overall deliverables that we aimed to provide, in order to fulfil the Project statement outlined in **Section 3.0**, are as follows:

1. An overview (mind-map) that categorises how the shop floor indicators can be utilised for analysis.

2. Inspect and propose the best-fit model explaining the relationship between the shop floor parameters and overhead costs per pair.

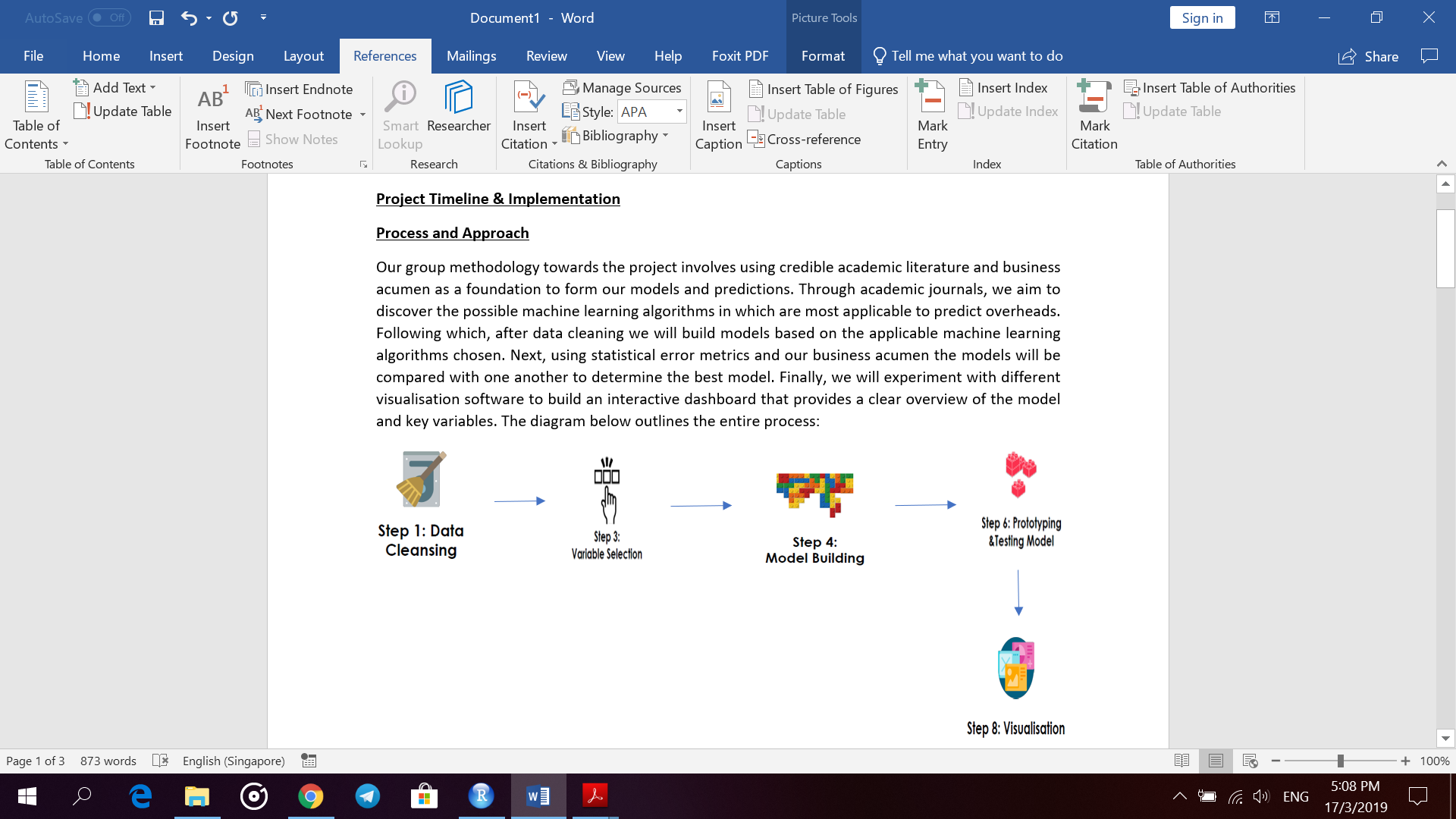
3. Dashboards displaying the eventual projected overheads in relation to the relevant shop floor indicators (e.g. what are the most significant parameters to be displayed, most suitable visual representations such as historical trends and what relevant software to utilize for visualization) Prototype of the dashboards, and test runs of the model using new user-inputs to find out expected overheads

4. Formal documentation of the standardised datasets, codes, formulas, and findings

**Project Timeline & Implementation**

**Process and Approach**

Our group methodology towards the project involves using credible academic literature and business acumen as a foundation to form our models and predictions. Through academic journals, we aim to discover the possible machine learning algorithms in which are most applicable to predict overheads. Following which, after data cleaning we will build models based on the applicable machine learning algorithms chosen. Next, using statistical error metrics and our business acumen the models will be compared with one another to determine the best model. Finally, we will experiment with different visualisation software to build an interactive dashboard that provides a clear overview of the model and key variables. The diagram below outlines the entire process:



**Data Pre-processing**

Data pre-processing involves the process of transforming raw data into consistent and useful data. Additionally, the cleaned data must be fitted into a schema that can be analysed easily. After discussions with StarAsia, our group came up with a schema that would fit the analysis desired – to identify which variables influence overhead cost per pair of shoes. Next, we sieved out all the possible variables related to overheads and collated them into the schema. The variables cleaned were chosen based on our business logic. For instance, variables such as EOLR effective were not chosen for cleaning due the fact that these were hypothetical numbers and had no predictive power towards overhead costs. The diagram below depicts variables chosen, and the structure of the schema.

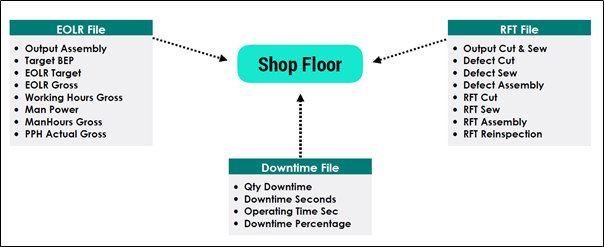


Figure X: Variables Cleaned (Appendix A)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Plant | Cell | Date | Variable 1 | Variable 2 | Variable 3 |

Figure X: Schema Structure (Sample of Data is in Appendix B)

**Exploratory Data Analysis**

After the data had been cleaned, our group conducted a brief exploratory data analysis on the cleaned data. The simple analysis was focused on plotting the variables in Figure X against overhead cost. From this, it would help us to gain a better understanding of the data and to choose the variables to be included in our model later on. Additionally, it would be able to help us observe the relationship between selected variables and overheads costs.

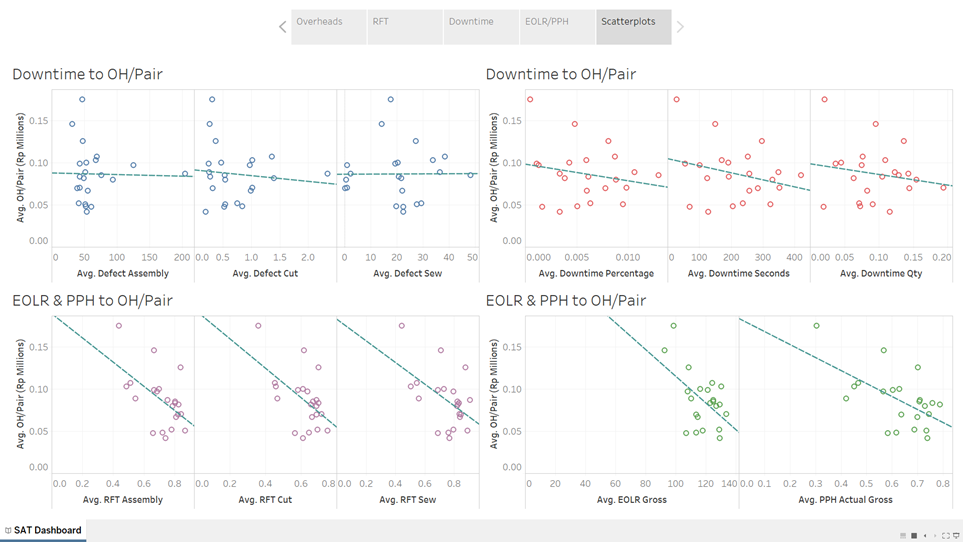


Figure: Sample of Exploratory Scatterplots

From the scatterplots plotted, we primary observed that most variables had no distinct linear relationship with overhead cost per pair of shoes at all. As such, our group decided that apart from running a simple regression model, we had to experiment with other algorithms as well.

**Model Development**

Through academic research and considerations of the deliverables at hand, our group chose three primary algorithms that we would experiment on.

**Linear Regression**

While it was mentioned earlier that linear regression might not be the best method to build our model, it does however provide an equation to that links overheads per pair to the variables. This was especially useful considering that it was the one of the deliverables expected. It also provides a numerical relationship between a variable and OH cost per pair of shoes. (E.g holding all else constant, how does 1 unit increase in Output Assembled affects overhead costs) Moreover, linear regression also helps to identify the significant variables relating to the dependent variable – overhead cost per pair of shoes. Given that we had to identify key factors that influence overhead cost per pair of shoes, this aspect was highly desired. Lastly, linear regression is a model that is simple to understand and is highly effective for predicting numerical figures such as overhead costs (Bruns & Caragnano, 2017).

**Lasso**

Another algorithm our group decided to experiment upon was Least Absolute Shrinkage and Selection operator, LASSO for short. It is a regression method that performs variable selection to improve prediction accuracy and interpretability of the end model it produces. Additionally, LASSO is especially helpful when we have large number of variables but a small number of observations (Fonti, 2017) which was one of the issues we faced. On top of that, LASSO also provides us with an equation linking overhead costs to the variables.

On the other hand, there are some limitations with LASSO. While LASSO aids us in variable selection, it does not determine the which variables are the most significant relative to one another (i.e. there is no ranking on variables). This may be a hindrance given that analysing and determining what drives overhead costs is a crucial component of our project. Nonetheless, in spite of this, our group proceeded on with this algorithm as it offers an a comprehensive method of selecting variables.

**Random Forest**

The third algorithm that we chose to build our model with is Random Forest. Random Forest is an ensemble learning method which is use for classification and regression. It operates by constructing a multitude of decision trees and combines them together to get a more stable and precise prediction (Donges, 2018). Manufacturing processes often produce large amounts of high-dimensional data in which machine learning algorithms like random forest are well-equipped to handle (Wuest, Weimer, Irgens, & Thoben, 2016). In addition, random forest is also able to rank the importance of the significant variables identified and this added feature would help provide a more detailed analysis of the results.

There are also limitations associated with the algorithm. The algorithm is based on decision trees and does not provide an equation. As a result, the model produced is not easily interpretable, and it would be difficult to gauge the numerical impact of each variable to overhead costs. Nevertheless, we decided to select this algorithm as it known to provide accurate results. Additionally, our group felt that using a classification model would better explain the data given its high dimensionality.

**Development of Visualisations**

The second phase of the project involved developing dashboards for use by SAT’s management. During this process, we had two primary objectives in mind when designing the visualisations.

Firstly, the dashboards should create a straightforward visual format for data inputs and provide a consolidation of offline updates to the data set. Currently, SAT’s data and visualisations (in the form of Excel Graphs) are dispersed amongst various Excel sheets, making it difficult for management to analyse the data easily. The dashboard will thus integrate data from various sources and helps to communicate this data in a comprehensive and coherent manner.

Secondly, the dashboards aim to provide a quick identification of trends and uncover discrepancies or outliers in the SAT’s data. This entails the descriptive level of analytics, whereby management will be able to make business decisions based on historical data. For example, the visualisations may highlight underperforming units in the business or factory-wide parameters that have declined over time.

With these objectives in mind, we compare the various visualisation softwares available on the market and suggest a recommended software for SAT’s needs. Using this software, we build a mock-up of the dashboard using SAT’s data, to demonstrate its functionalities.

**3.0 Model Development**

**Data Preparation**

While the data for shop floor parameters was recorded in a daily format sorted by the respective factory plants, our dependent variable (overhead cost) was recorded according to months and was representative of the factory as a whole. (Figure X)

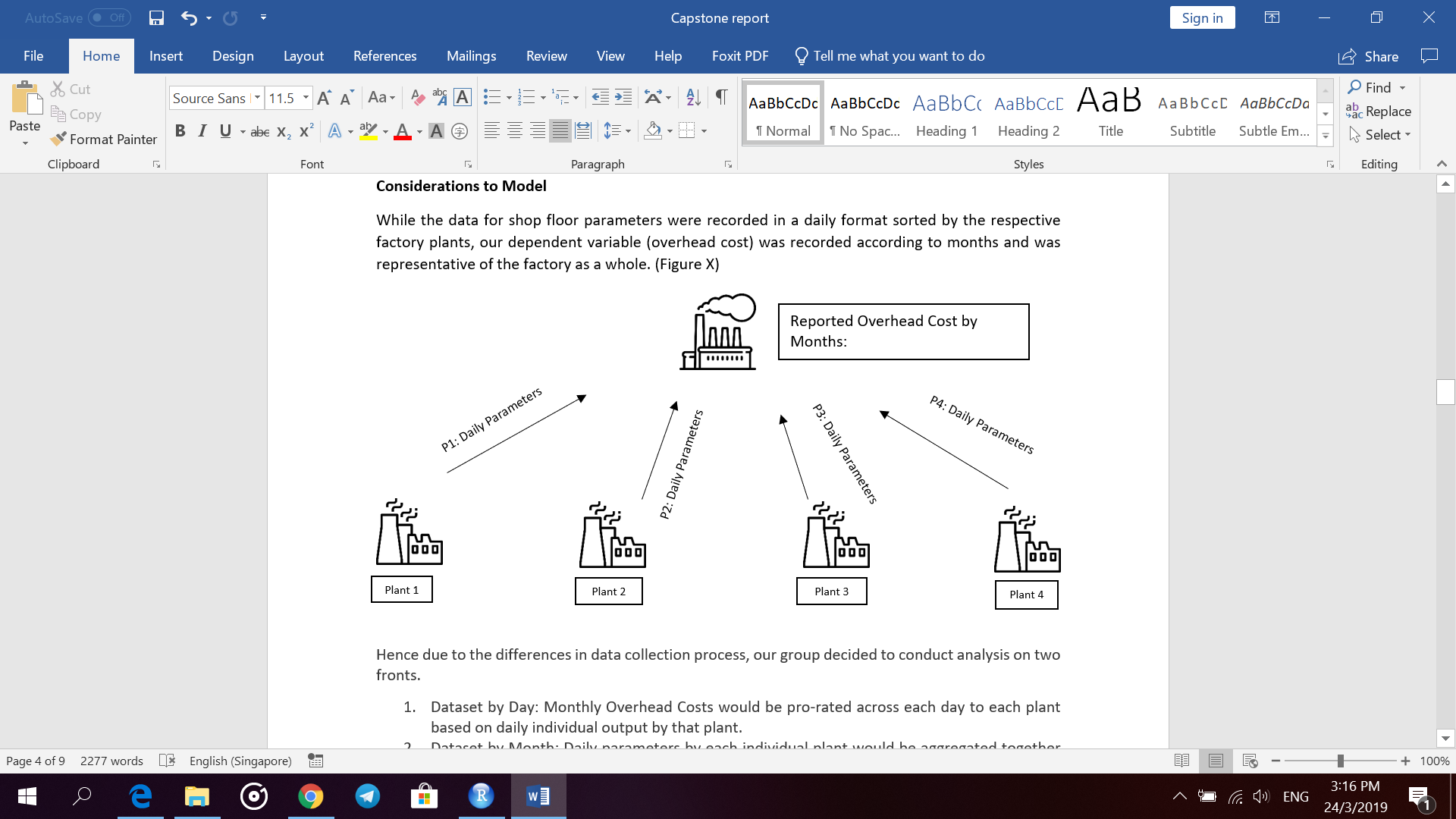
****

Figure X: Data Collection Process of SAT

Hence due to the differences in data collection process, our group decided to conduct analysis on two fronts.

1. Dataset by Day: Monthly overhead costs is divided by the number of days in that month. Overhead costs is assumed to be the same each day (E.g. in Appendix C).
2. Dataset by Month: Daily parameters by each individual plant would be aggregated together by months ( Appendix D).

For both approaches, days where there was zero output was removed from the dataset completely as it was assumed that there were no overhead costs incurred when the factory is not operating. The dataset was then split respectively into training (Jan17- May18) and testing (Jun18 – Oct18). Next, all three algorithms – linear regression, LASSO & Random Forest were used to train the model on the training dataset. Subsequently, the trained model would then be tested for its predictive accuracy based on the testing dataset. To compare all the models, Mean Absolute Error (MAE) was selected. MAE was selected as all individual differences between predicted and actual value are assumed to have equal weight.

**Selection of Algorithm**

There were a number of factors that we used to compare the different algorithms. Firstly, as mentioned earlier, as a general rule, the model that has the least MAE should be chosen as the best model. However, we also know from the visualizations that the overhead costs do not have a linear relationship with the independent variables. Hence, a linear regression model might not be suitable even though it returns the least MAE in some cases. Similarly, for LASSO, it tends to overfit the model thus adds unnecessary volatility. As such, moving forward, the group feels that Random forest will be the most appropriate algorithm due to its accuracy as well as its capability to handle a dataset with high dimensionality. The most accurate model (Random Forest) will be presented in the main report, while other models that the group have tested (Linear Regression, LASSO) will be attached in the Appendix for comparison purposes. The group eventually felt that Random Forest was the most appropriate model to use because of its accuracy and ability to handle large datasets with high dimensionality. However, Star Asia Trading can look at the Appendix for the linear regression/LASSO models if they want to determine the impact on total overheads with a manual input of a variable because of the availability of coefficients.

**Most Accurate Model Results**

As mentioned above, this section focuses on the interpretation of the best model from both approaches.The results of the different models for both daily and monthly approaches can be found in Appendix D.

Out of the three algorithms run, the models based on the Random Forest algorithm had the best predictive value for both approaches(daily & monthly). After testing it on the test the daily model had an MAE of 759.9447 rupiah (SGD 0.072). Likewise, the monthly model had an MAE of 27030.47 rupiah (2.57). Additionally, the following significant variables were identified by the algorithm.

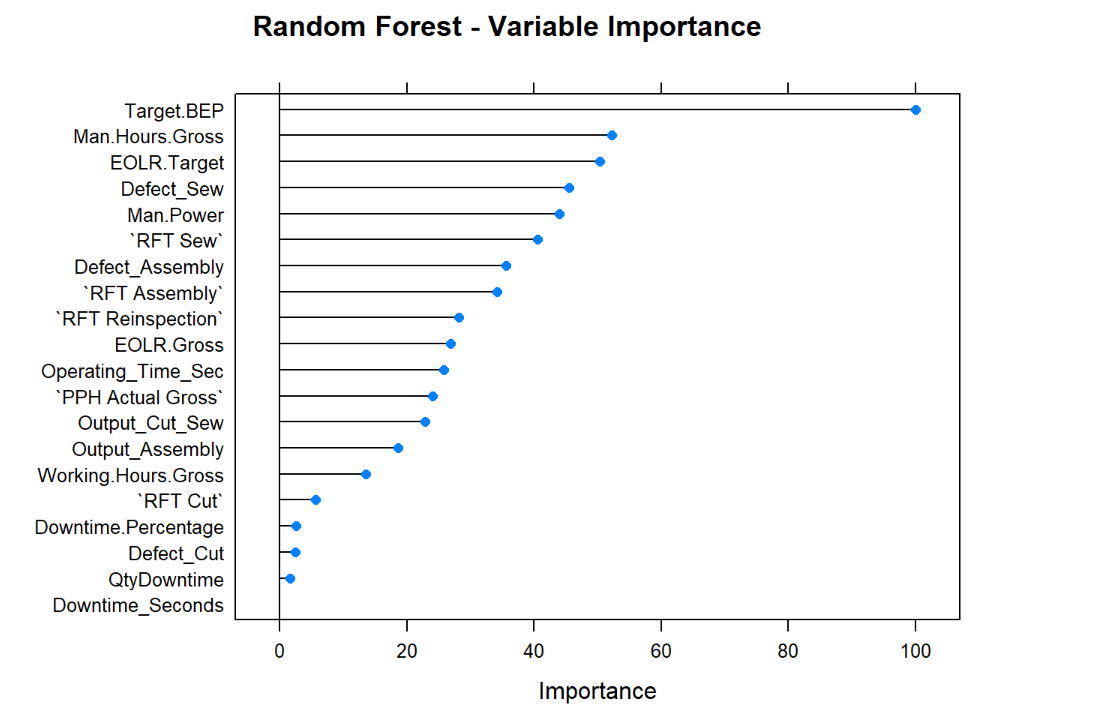


Figure X: Significant Variables Identified using Daily Approach

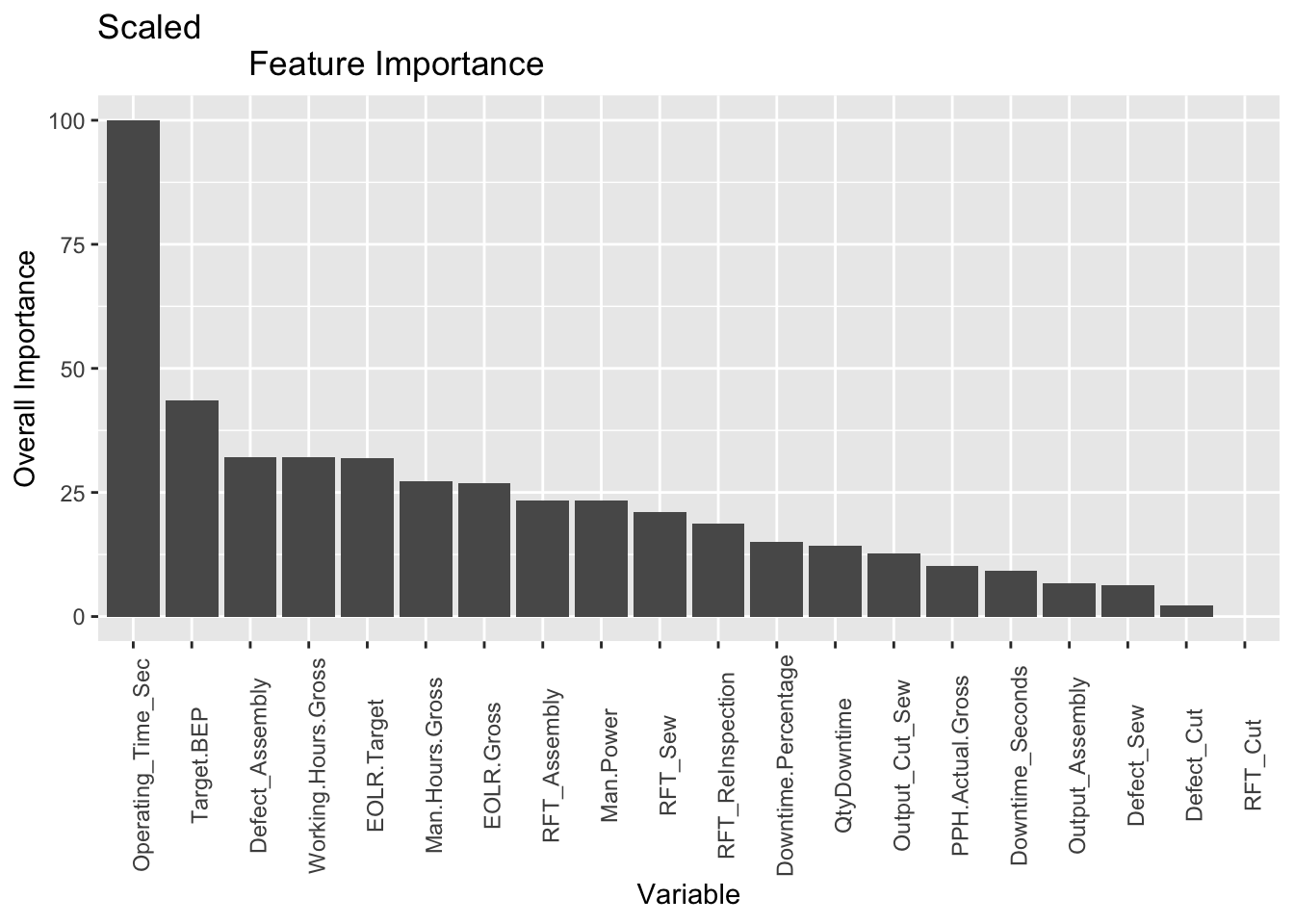


Figure X: Significant Variables identified using Monthly denominations

**How Random Forest determines Variable Significance**

For numerical predictions using random forest, variables are deemed significant based on its predictive value. It primarily does this in two ways statistically.

1. Percentage increase in mean square error (MSE) to accuracy and the variable is permuted (IncMSE). For each feature, the algorithm will randomly permute values of the feature and subsequently measure the resulting increase in error. The higher the increase in error, the more the predictive power of that variable.
2. Gini Impurity. Gini impurity is a metric used in decision trees to determine which variable to split and at what threshold to split the data into smaller groups. The higher the purity the better.

**Group’s Opinion on Significant Variables**

**Similar Significant Variables**

As seen both the daily and monthly approach determined that Defect\_Assembly & Man Hour Gross are ranked highly in accordance to variable importance. The more defects the factory produces, the greater the overhead costs incurred. This is reasonable considering that more overheads such as utilities, small tools etc will be used to rectify the defects. Similarly, as man hour gross increases, more overheads will be incurred which drives up overhead cost per pair of shoes.

On the other hand, an interesting factor to note is that target variables such as Target BEP, EOLR target are deemed to have predictive value for overhead costs under both approaches. This is abnormal as these variables do not contribute any quantifiable costs to the manufacturing process since they are simply man-made estimates. A possible explanation is that this is probably due to the inherent limitations of our model. As we are predicting total overhead costs (fixed and variable), variables which had consistent numerical values were effective in predicting fixed overhead costs. Thus, target variables were deemed to have strong predictive values.

**Different Significant Values**

Expectedly, there were some differences in the significant variables between the daily and monthly approaches. For instance, Operating\_Time is lowly ranked under daily while under monthly it was ranked highly. From an accounting perspective, operating time should have a strong predictive power given that the longer the factory operates, the more overhead costs will be incurred. However, this is not reflected in the daily approach. A possible explanation could be that there is a limited number of datapoints in our monthly dataset. For our daily dataset we have a total of 4444 data points (3636 for training, 808 for testing) while for monthly we only have 22 points (17 for training, 5 for testing). As a result, under the monthly model operating time differs substantially from month to month.The algorithm might have felt it had a high predictive value compared to the daily model where differences between operating time per day is relatively small.

In sum, while both approaches has their merits, the group feels that the monthly approach would be more applicable to SAT. A major flaw in the daily approach would be that daily overhead costs are assumed to be constant each day regardless of output. Thus, due to this estimation, it creates further volatility in the model produced. Hence, moving forward, analysis on based on the monthly approach would be more appropriate given the manner the data is obtained.

**Model Development**

**Limitations of model and suggestions for improvement**

According to the problem statement, Star Asia Trading wanted to understand the relationship between shop floor parameters and overall overhead costs. After doing the initial analysis using overall overhead costs as our predicted variable, our group feels that this method of analysis has a limitation given our domain knowledge about accounting. In management accounting, total overhead costs can be split into both fixed and variable overhead costs. Since the shop floor parameters relate mostly/entirely to the output of the firm, it will not be logical to use shop floor parameters to predict the fixed portion of overhead costs because Star Asia Trading might have invested in fixed assets or rent contracts during the year that is independent of production. This confounds the relationship between shop floor parameters and total overhead costs.

A better suggestion that our group would like to bring up to better predict total overhead costs is for Star Asia Trading to determine its fixed portion of total overhead costs based on existing rent contracts and expenses relating to fixed assets like machinery. Our group can instead predict the variable portion of overhead costs using shop floor parameters, and this approach will be more logical because there is a stronger and non-confounded relationship between these factors.

Another limitation of this model stems from the design of the database. In order to have a database design that is easy to understand, we agreed with SAT to design the schema as per Figure X: Schema Structure located above. However, the limitation of placing the variables in such a structure means that when one variable (for e.g. Output\_Assembly) is found to be important in the model, we have to assume that this variable is important as a whole regardless of which plant or cell it belongs to. However, this might not be the case for Star Asia Trading, because certain plant/cells have different specifications and variables should not be aggregated regardless of plant or cell. One suggestion for a better analysis is to change the schema design so that non-aggregated measures can be used. The new schema design should look like this:

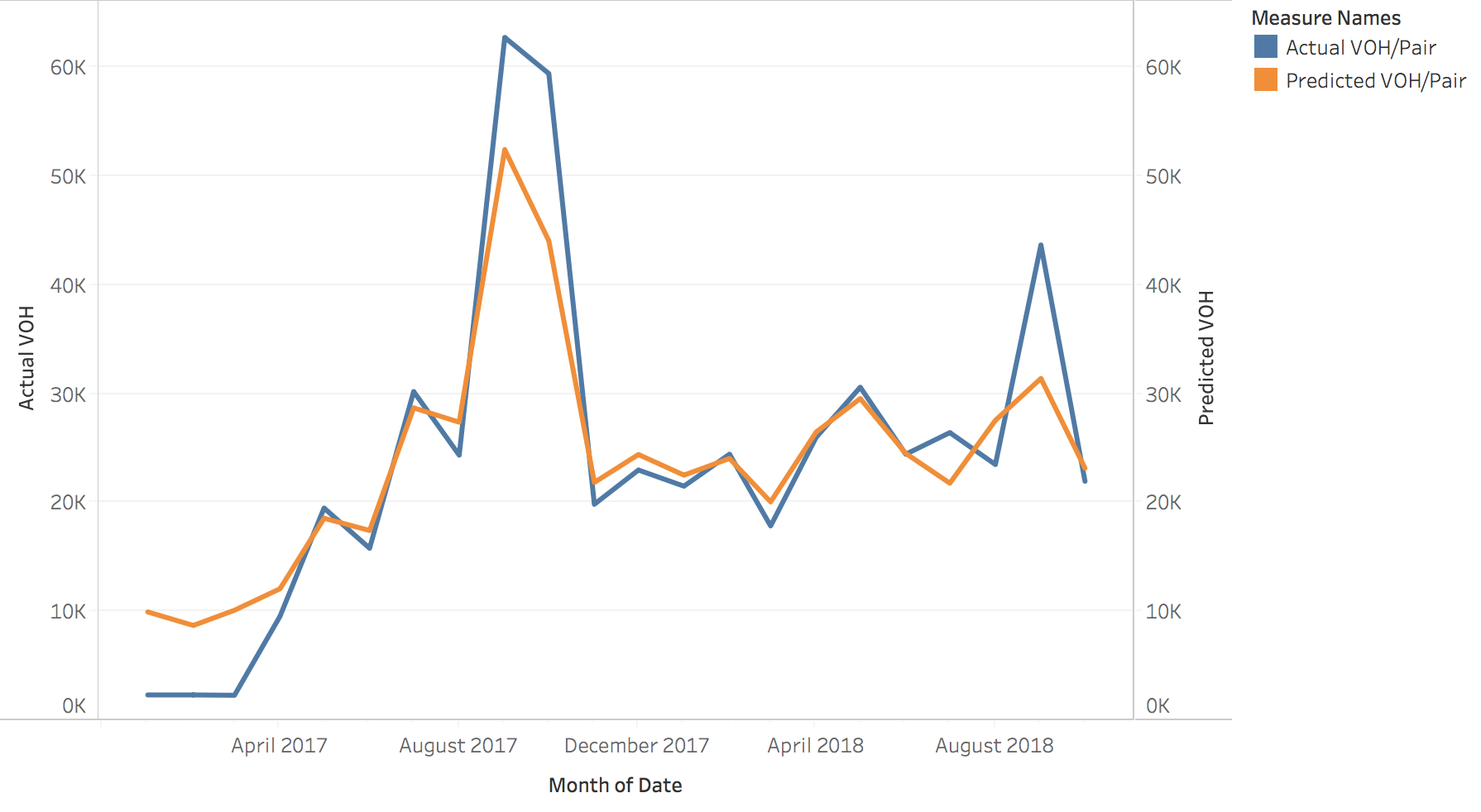
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | P1C1\_Variable 1 | P1C1\_Variable 2 | P1C1\_Variable 3 | P1C1\_Variable 4 |

This way, it allows the model to take in the variables per plant and cell, which enables the analysis of important variables for each particular function of its manufacturing process.

**Improved Model Generation**

Predicting variable overhead costs

Another approach to the client’s problem is to predict only the variable portion of overhead costs using shop floor parameters, as mentioned earlier.



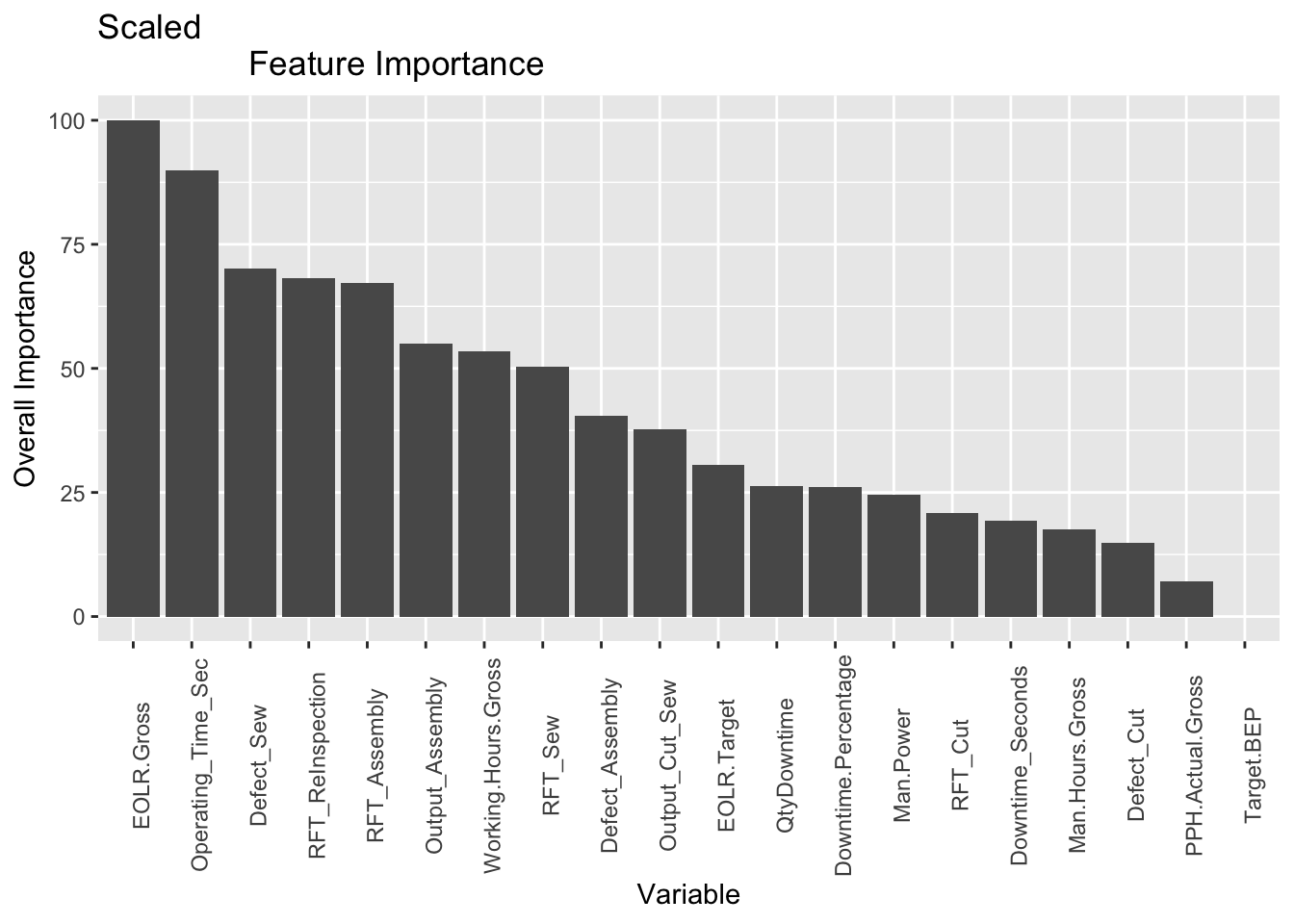


Figure X: Significant Variables using Monthly Denominations

MAE: 4448.603 rupiah ~ $0.42 SGD variable overheads per pair

Useful insights can be gained from the graph showing actual versus predicted variable costs per pair (VOH/Pair). For instance, there is especially high variance when actual VOH/Pair is extremely low or high when compared to predicted results. This might be one area to look into for the client because the behaviour of VOH/Pair is expected to be relatively constant over time. It might be due to significant events that happened during Jan - March 2017 and September 2017/2018 that resulted in the extreme highs and lows.

The variable importance results seem to make much more sense now. In this model, the target variables (i.e. Target BEP and EOLR.Target) are way less useful in predicting variable costs, which matches the team’s mental model that targets shouldn’t be useful in predicting overhead costs.

The recommended purpose of this model is for prediction. This is due to the model having the best estimate of VOH/Pair out of all models. SAT can use this model to predict the variable portion of overhead costs, and add it to the fixed costs from its longer-term contracts/fixed assets to get a prediction of total overhead costs. On the other hand, the variable importance generated by this model does not appear to be helpful to SAT because it does not yet specify which variable is important in a particular plant and cell division. The client will not be able to use this model to cut down on overhead costs accurately. This limitation is addressed by the second new model below.

Predicting variable overhead costs using non-aggregated schema

Another approach that SAT had mentioned during week 9’s presentation was to have an additional model that does not aggregate the variables into a firm-wide variables. This relates to the schema design and was discussed as a limitation of the original above. The team has come up with a refined schema and was able to conduct an analysis with the new schema.





Figure X: Significant variables using non-aggregated data

MAE: 7238.86 rupiah ~ $0.69 SGD variable overheads per pair

Similarly, we have the actual versus predicted VOH/Pair and variable importance results for this second new model. However, the recommended purpose of this model is to identify the variables specific to each plant and cell in order for the client to optimise.

From this variable importance plot SAT can view the variables in non-aggregated form and important variables include Plant 2 Cell 2’s EOLR Gross, Plant 1 Cell 1’s Man Hours Gross etc. This is much more useful for the client because they will know which area of their shop floor targeted at the plant/cell to optimise. However, the limitation of using the random forest algorithm is that there is no coefficients available for the client to work on. This means that the client has to rely on either business sense or the coefficients from the linear regression in order to make a decision to reduce or increase variables.

This model is less accurate at prediction as seen by the higher MAE and the reason might be due to its non-aggregated nature, which tends to perform worse than aggregated models.

**Justification of usefulness of models towards achieving objectives**

Rationale behind choice of model

The group found that there is no linear relationship between most of the shop floor parameters and total overhead costs, which reduced the effectiveness of using a linear regression model. The group eventually went with Random Forest out of the two remaining models (Random Forest & LASSO) because LASSO attempts to shrink the coefficients, which might unnecessarily reduce the volatility in overheads. This might not be what Star Asia Trading wants in a model and therefore, the Random Forest model is used. However, the limitation is that there is no coefficients involved and STA will not be able to instantly know the changes in overhead costs given an input without running the Random Forest model for predictions.

In terms of assisting the client to find a relationship between shop floor parameters and total overhead costs, all three models that the group has done up will allow the client to better understand their business. The models are comprehensive too because it allows the client to do different levels of analysis on their business according to their decisions. For instance, they can use the model that predicts the variable portion of overhead costs if they agree that only variable overhead costs have a relationship with shop floor parameters. If they do not agree with this logic, then they are still able to revert to the very first model(Section 3.X) where total overhead costs are predicted. Ultimately, it is up to SAT to select the model that best fits their their objective.

**Visualisation Outputs**

Choice of Software

In this section, we provide a comparison of different visualisation softwares that are available on the market. The recommended software will be based on how well it fits the needs of SAT, as well as the practicalities involved in building the mock-up for the project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Tableau** | **Microstrategy** | **Microsoft Power BI** | **Qlik Sense** |
| **Usability** | Intitutive drag-and-drop and attractive interface  Supported by mobile platforms  Highly flexible, allows for “overlapping” which increase screen space ergonomics | Intuitive drag-and-drop  Interface  Supported by mobile platforms  Report creation function  Lacks predictive of prescriptive tools, and certain scientific visualisations | Intuitive for users with experience with Microsoft software  Performance issues with large datasets | Flexible, as it allows users to tweak many aspects in the visualisations  Dashboards and reports are easy to navigate, but building reports require proficiency in SQL and Qlik’s Proprietary language |
| **Learning Curve** | Low  Vast library of video learning materials | Moderate | Low  Familiar  interface for Windows users | Moderate |
| **Support for Data Source types** | 40 + data sources, including file formats, data systems and cloud systems | 50+ database/data source platforms | Supports most kinds of data sources | 40 + data sources |
| **Connectability to Cloud Databases** | Yes | Yes | Yes | Yes |
| **Differentiating factor** | Features  “Data blending”, which allows for collaboration in real time  Rich library of visualisations that are not available on other platforms  Powerful community collaboration | Provides a “Microstrategy ecosystem” that integrates databases and visualisation tools | Affordable option and offers a free version  Integration with MS products | Recognises relationship between data items without explicit preconfiguration |
| **Free Account** | Free public version | 30-Day Free Trial | Free version: Power BI Desktop | Free personal version |
| **Pricing** | Ranges from USD35-70 per month | By quote only | Power BI Pro: USD9.99/month per user | Qlik Sense Enterprise: By quote only |

*Table X: Comparison of Various Visualisation Tools*

Comparing the various visualisation softwares, Tableau appears to be the most appropriate for SAT’s requirements. Firstly, Tableau allows for intuitive and easy dashboard creation for users as opposed to the other softwares. For example, Qlik Sense and Microstrategy may require some form of proficiency in proprietary computing languages (Bobriakov, 2018). As such, Tableau allows the user to create useful dashboards without much prior knowledge or training.

Tableau offers a range of products at competitive prices. While Power BI Pro is the most affordable option, performance issues with large datasets may reduce the effectiveness of dashboards created using the software (Bobriakov, 2018). Since SAT aims to build a smart-factory which involves vast amounts of live data, Power BI Pro may not be suitable for their needs.

Meanwhile, Microstrategy’s professional versions are priced at a by-quote-only basis (Pardo-Bunte, 2018). For Qlik products, users have experienced complex pricing schemes that result in greater-than-expected fees (Mazenko, 2016). As such, Tableau provides a transparent and relatively moderate price for better dashboard performance.

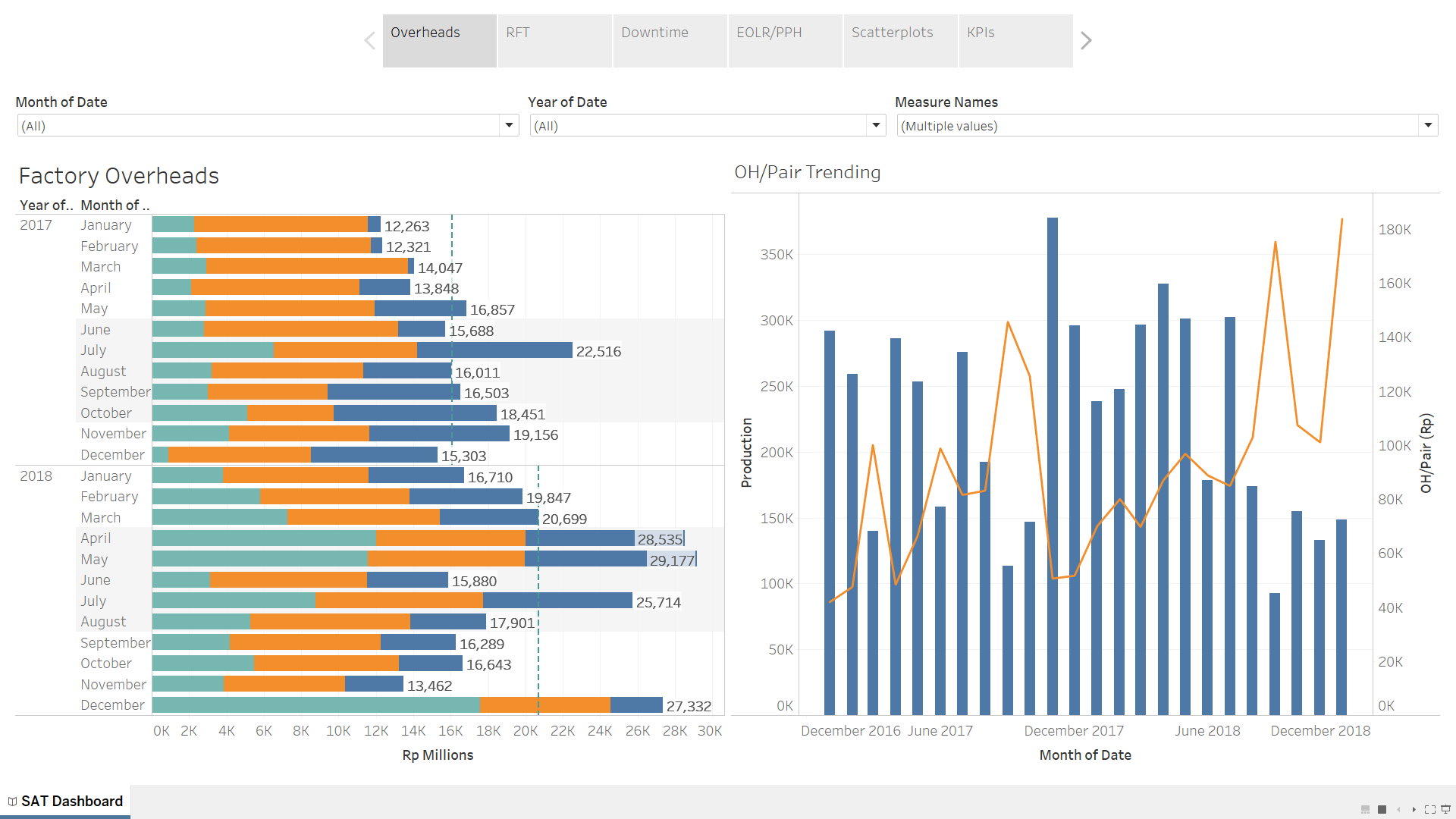
In addition, Tableau has one of the largest community of over 57,000 users (Marr, 2018), creating an online network which provides a vast support for users. Complex queries can be addressed through easily-accessible educational videos and online forums.

Given these factors, we have created the visualisations using Tableau and SAT’s historical data. However, it is noted that most visualisation platforms can create dashboards that have comparable functions. As such, the dashboards will primarily be used to illustrate the usefulness of visualisations, and SAT ultimately has the discretion to select the software that they deem most suitable for their needs.

SAT Overheads Dashboard

The following section documents the dashboards that have been created and their functionalities. The visualisations provide insights at the “descriptive analytics” level, with the intention to assist management in decision-making. In order to improve user-friendliness and navigation, we have stitched the dashboards into a consolidated story format.

Overheads Dashboard



*Figure X: Overheads Dashboard*

The first dashboard is provides a historical analysis of factory overheads data (Figure X). The top column (circled in red) contain three drop-down boxes, which allows the user to filter data according to certain time periods (Month or Year) and the type of overhead expenditures.

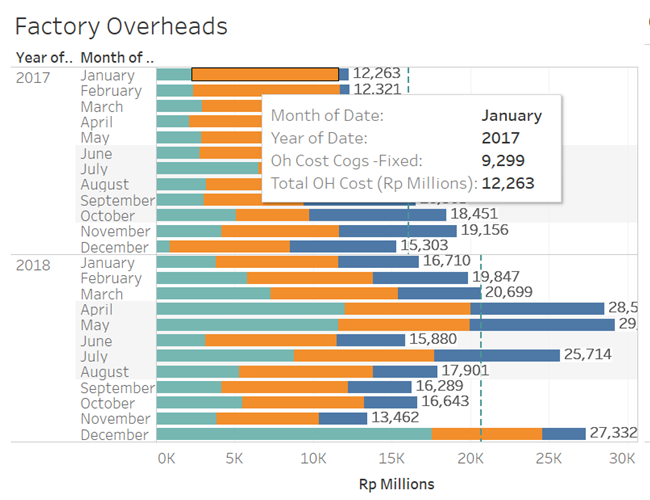


Figure X: Factory Overheads

The “Factory Overheads” dashboard provides a representation of overhead expenditures per month using a stacked bar chart, where the expenditure types are categorised as: Fixed Overhead (OH) Costs (orange), Variable OH Costs (dark blue) and Operating Expenditures (teal). Additionally, the tooltip (Figure X) provides information on the expenditure type and amount. The Total OH cost per month in Rp millions is presented at the end of each bar, along with the average Total OH cost for the year (represented with a teal dotted line).

This dashboard aims to give management a quick overview of overhead expenditures by category over the 2-year period. For example, actual figures can be easily compared with budgeted figures to identify the periods whereby the budget was exceeded. This helps management to decide if there is a need to curb activities where the factory has over-expended and to anticipate rising overheads for the coming months, based on seasonality.

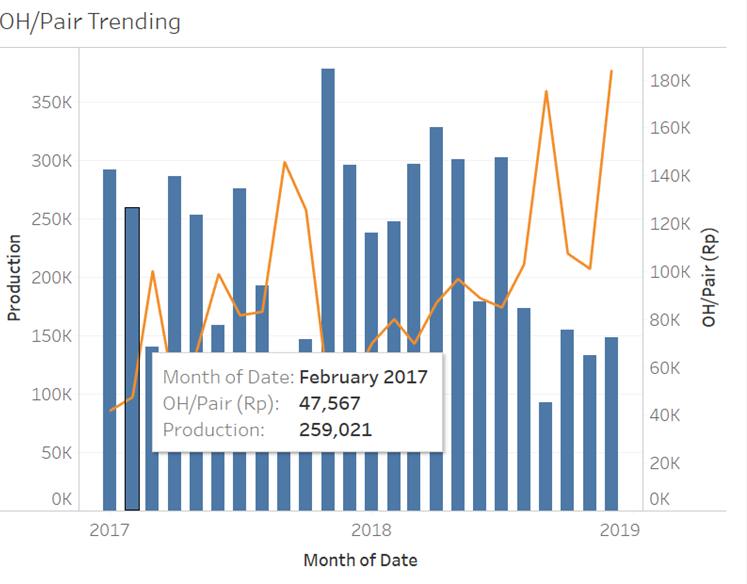


Figure X: OH/Pair Trending

Next, the “OH/Pair Trending” represent Average OH costs per pair (orange line) on a time series and Total Production per month (dark blue bars) during 2017 and 2018 (Figure X). This visualisation aims to provide a trending of key factory measures during the two-year period, allowing management to monitor them easily. Here, we observe that the Average OH costs per pair is experiencing an upward trend, which may require management’s attention and prompt an enquiry to investigate the underlying reasons.

Right-First-Time (RFT) Dashboard

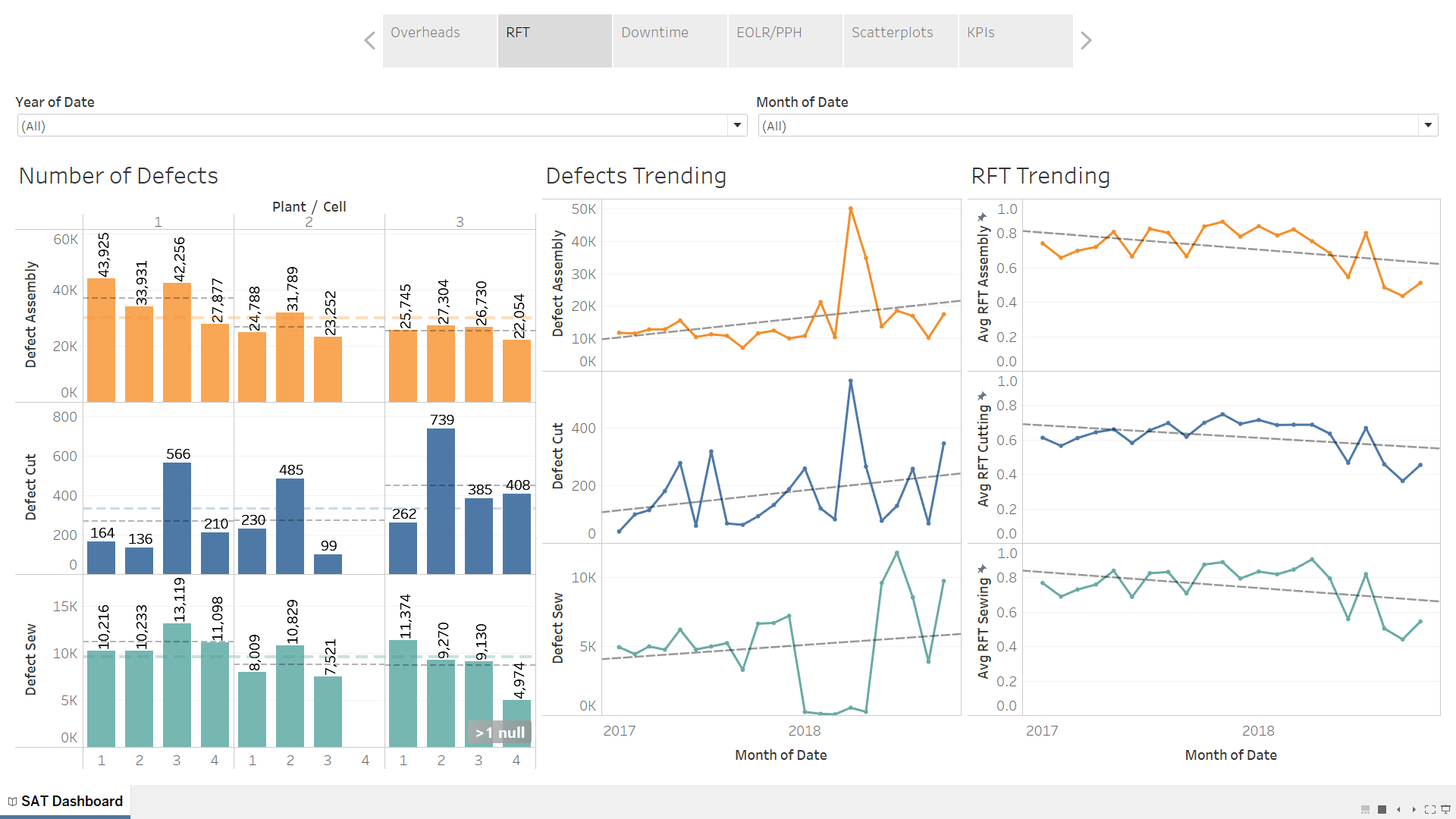


Figure X: RFT Dashboard

RFT is an important metric used by SAT’s management to measure the number of pairs that are produced without defects on the first attempt. The RFT dashboard thus provides an overview of defects and RFT within the factory (Figure X).

On the left panel, “Number of Defects”, bar graphs are used to visualise the sum each type of defects (Assembly, Cutting and Sewing), categorised into plant and cell. This enables management to quickly analyse the plants/cells that are generating more defects than the factory average. By identifying the highest contributor to defects, management can adopt a targeted approach towards improving the overall quality of production in the factory.

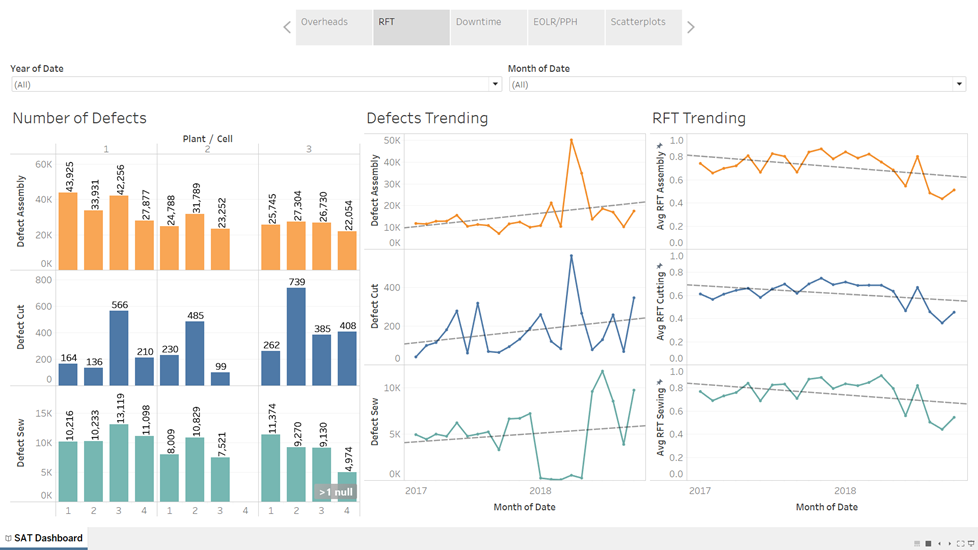
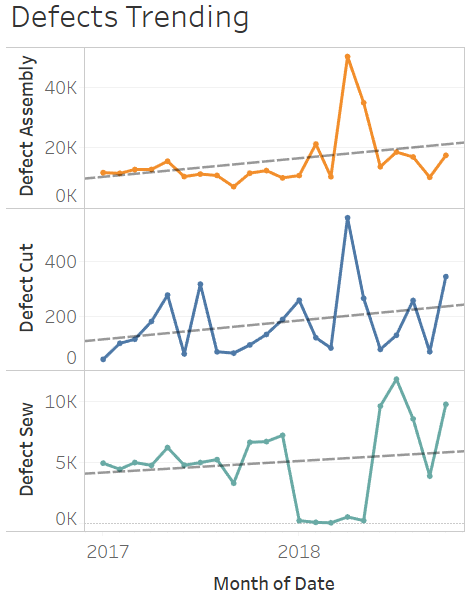


Figure X: Defects and RFT Trending

The remaining two panels, “Defects Trending” and “RFT Trending” shows the trends in defects and RFT respectively, which is similarly categorised into the defect types (Assembly, Cutting and Sewing). This allows management to conduct a time series analysis on defect measures - for example, RFT seems to be decreasing on all fronts during the two-year period and should be a concern for factory managers as it indicates increased wastage in manufacturing.

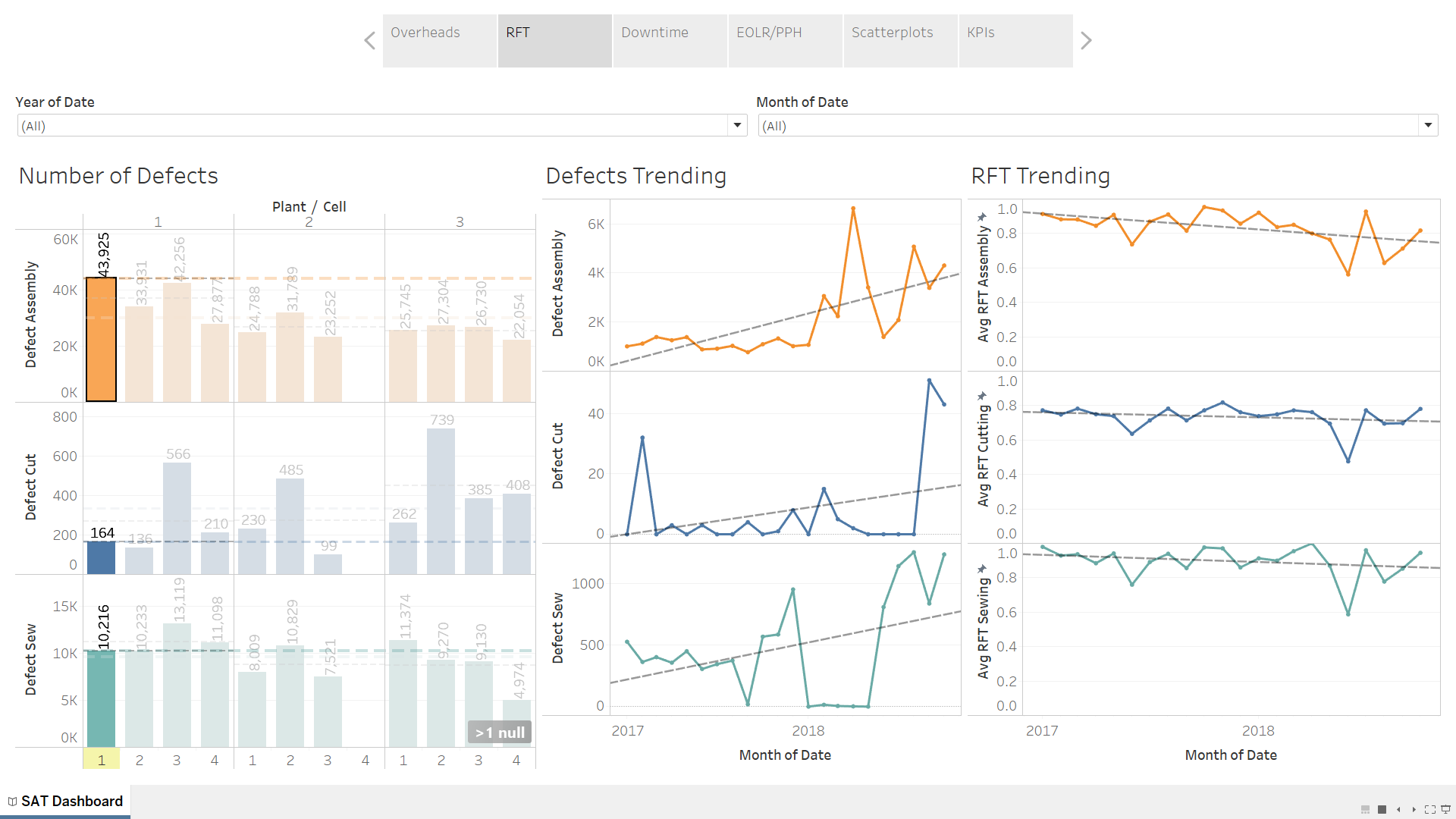


Figure X: Filtering the RFT Dashboard

Finally, users may filter data using the dashboard “Number of Defects” which is linked to the “Trending” dashboards, allowing for a more dynamic and granular analysis of specific plants/cells. In Figure X, the data in the “Trending” dashboards has been adjusted according to the selected plant/cell (Plant 1 Cell 1 in this case).

Downtime Dashboard

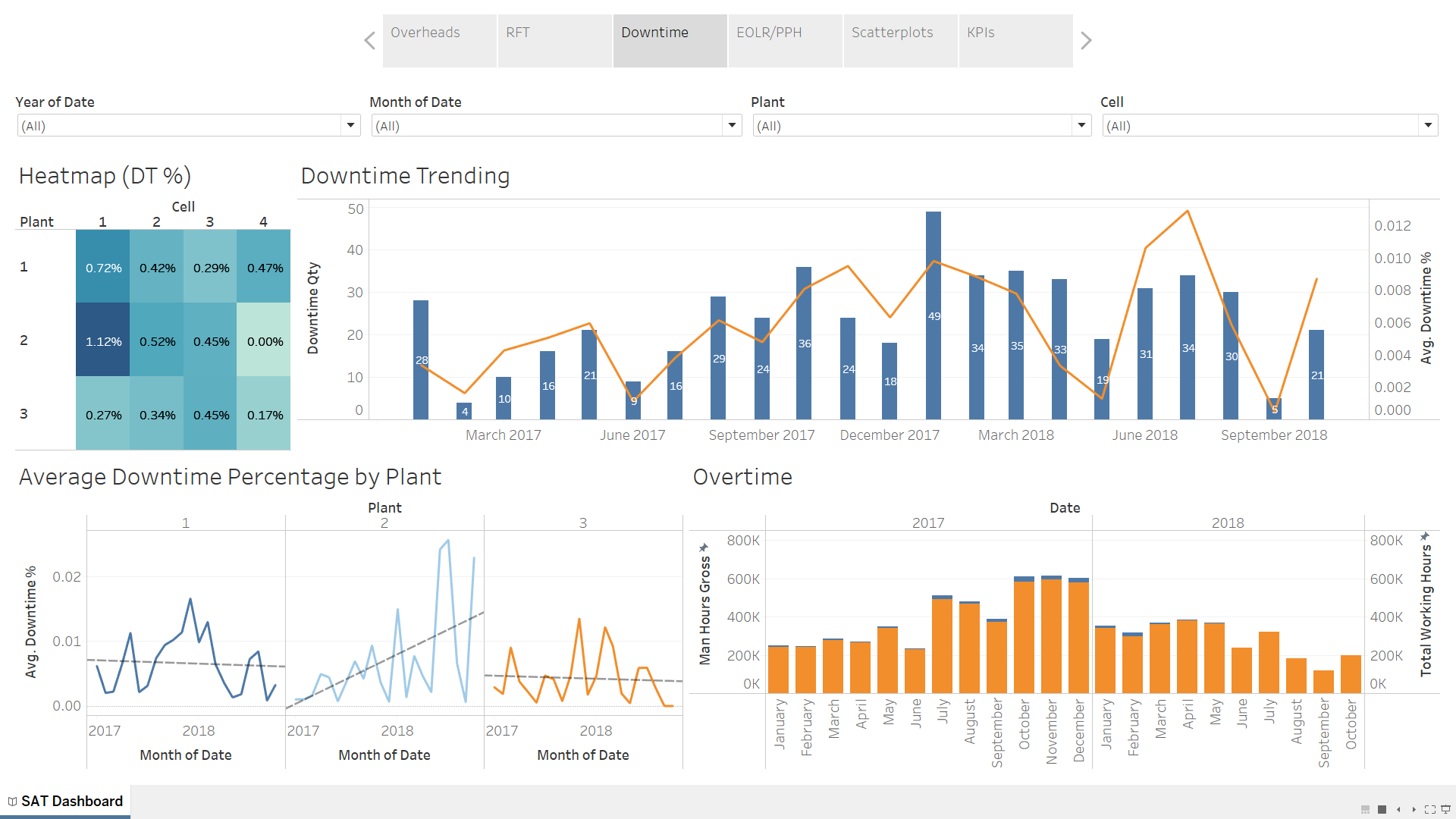


Figure X: Downtime Dashboard

The third dashboard focuses on Downtime metrics (Figure X). Firstly, we have a heatmap, “Heatmap (DT %)”, depicting Downtime Percentage during the period (measured as Downtime\_Seconds/Operating Time\_Secs) on the top left. Darker coloured squares accentuate the Plant/Cells that are experiencing greater Downtime Percentages, visually highlighting to management the business units that require additional attention. For example, management can enquire why Plant 2 Cell 1 is experiencing more than twice the downtime of all other units in the factory.

On the top right panel, “Downtime Trending” provides a time series for Downtime Quantity (dark blue bars) and Downtime Percentage (orange). This trending analysis supplements the information in the heatmap, which only captures the average Downtime Percentage for the entire period. If management notices that Downtime Quantity/Percentage has been increasing, they can launch an investigation to address the problem. Additionally, the trending analysis allows them to visualise the success of new initiatives to reduce downtime (if any).

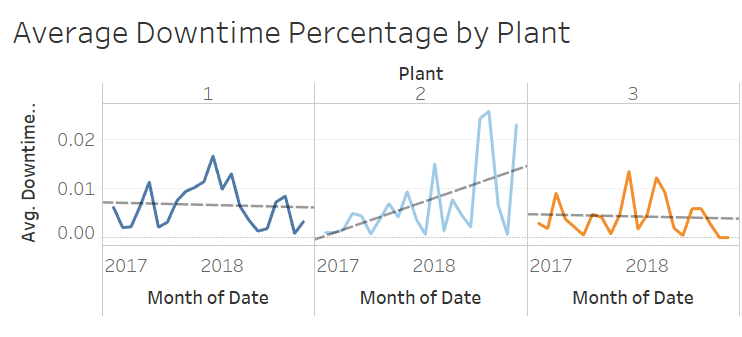


Figure X: Average Downtime Percentage by Plant

The bottom left panel, “Average Downtime Percentage Per Plant” allows for a more detailed breakdown by Plant (Figure X). Here, while Plant 1 and 3 have maintained a relatively constant Downtime Percentage between 2017 and 2018, Plant 2 has encountered a notable increase in average Downtime Percentage. Management can launch an enquiry as to why this is the case for Plant 2. As such, the dashboard helps to identify areas of the factory that are facing issues.

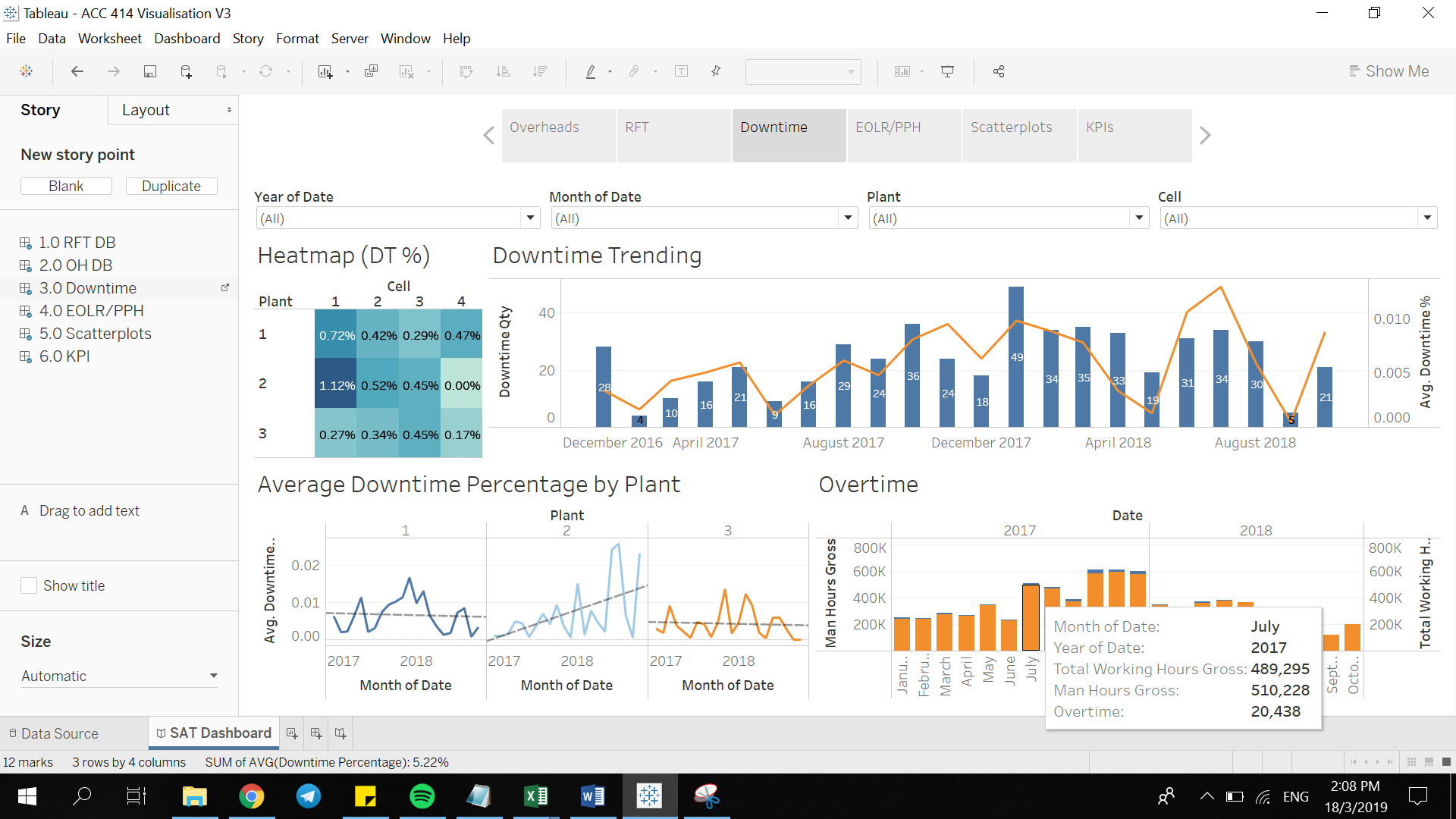


Figure X: Overtime

Lastly the “Overtime” panel depicts the amount of Man Hour Gross and Working Hours Gross. The excess Total Working Hours (blue bars) depict the the amount of Overtime for each month, with a tooltip providing numerical figures for each variable. Since overtime wages lead to increased overhead expenditures, this dashboard helps management to track the amount of overtime occurring at the factory.

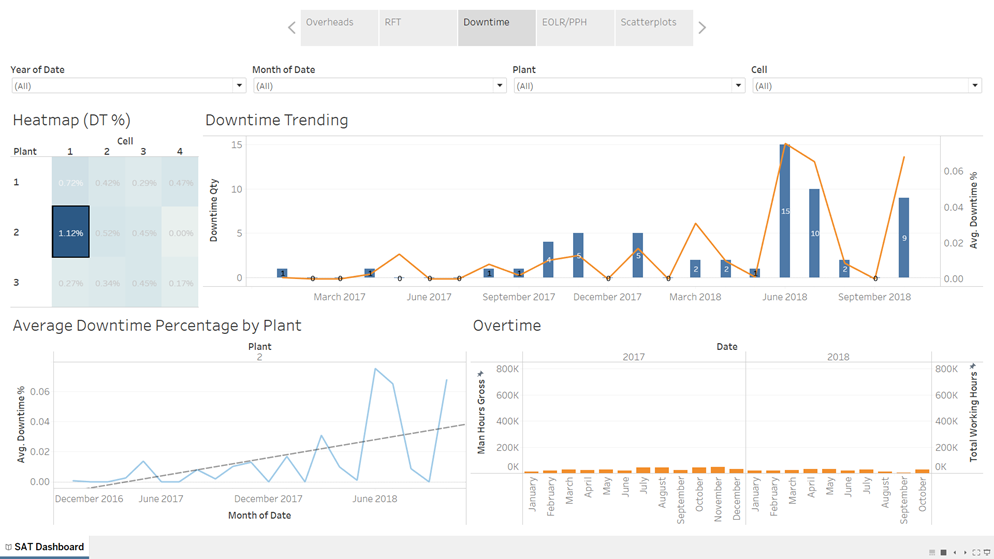


Figure X: Filtering the Downtime Dashboard

Similarly, by clicking a Plant/Cell on “Heatmap (DT %)”, data in the remaining visualisations in the Downtime Dashboard will be filtered accordingly. In Figure X, Plant 2 Cell 1 is selected. Alternatively, users may filter the dashboard using the drop-down boxes located at the top panels (circled in red).

End of Line Rate (EOLR) & Pairs per Hour (PPH) Dashboard



Figure X: EOLR & PPH Dashboard

EOLR and PPH are metrics used by SAT to measure the efficiency of the factory. “EOLR % of Target” and “PPH % of Target” (Figure X), provide a visual representation of actual EOLR and PPH (in darker colours) against targets set by management (in lighter colours). Additionally, the data is categorised by Plant and Cell, giving management a clear view of how individual Cells and Plants are performing.

The bottom panels, “EOLR Trending” and “PPH Trending” provide a time series of EOLR and PPH respectively. This enables management to identify persisting trends in the EOLR and PPH. Here, we observe that while EOLR remains relatively constant, PPH has declined significantly, indicating that the overall efficiency of the factory has decreased over time.

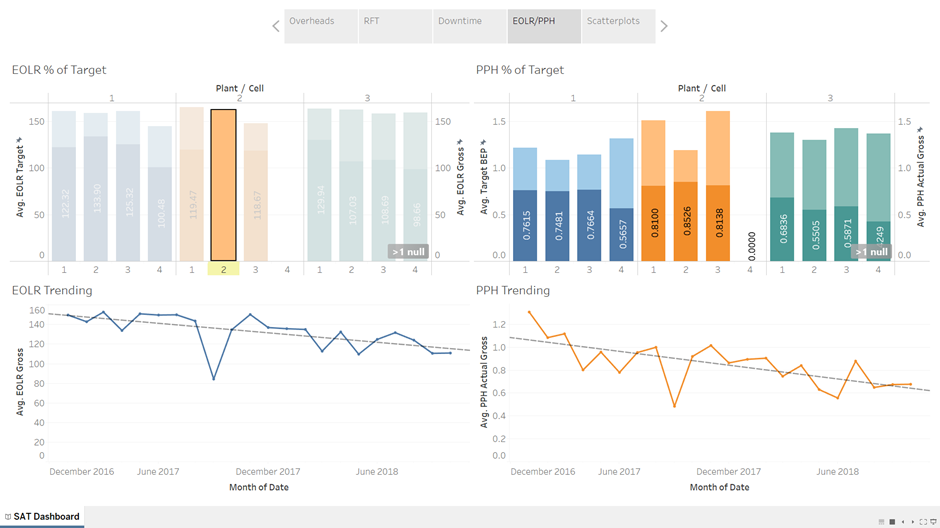
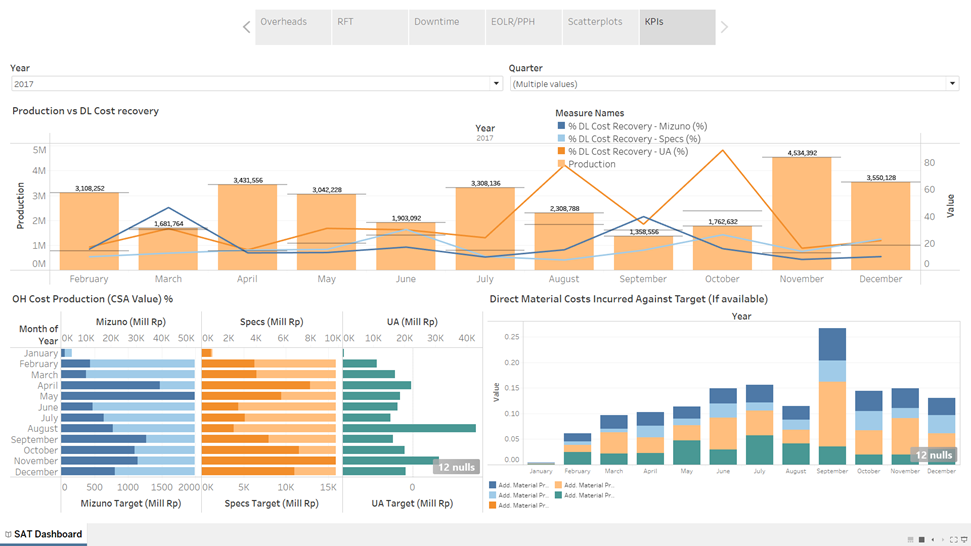


Figure X: Filtering the EOLR & PPH Dashboard

The dashboard can also be filtered in a similar fashion as the previous two dashboards - in Figure X, Plant 2 Cell 2 is selected, which applies a filter on the bottom panels.

KPIs Dashboard



* + How we organized data on visualizations
    - For each visualization
  + Screenshots - Walkthrough of your demonstration using screen captures (if any)

**Further Recommendations**

**Going Forward**

Although the main focus of this project is about the prediction of overheads, it is a stepping stone towards the bigger picture of SAT’s transition towards becoming a smart factory. As such, while the group focus towards the project is about the accuracy of the predictions, the group has also taken the initiative to propose certain measures that may be useful to SAT in the long run.

**Artificial Neural Networks (ANN)**

While the group has suggested various models we also acknowledge that there are other ways which may be better in addressing the problem statement at hand. In the course of our research and discussion with SAT, the group was recommended by Tian Tian (SAT’s data scientist) other machine-learning algorithms that may be more applicable but due to the tight schedule, we were unable to effectively apply it. One of which was the possibility of using Neural Networks.

Neural Networks are computing systems that mimic the way in which biological neural networks process and handle information. Such systems able to perform tasks without being explicitly programmed to follow any task-specific rules. In a manufacturing context, ANN have a satisfactory prediction ability often better than regression methods as suggested by the group (Setyawati, Sahirman, & Robert, 2002). In particular, long-term short term (LSTM) neural network might be useful to SAT to achieve their objectives. LSTM is highly dependent on past data, using it to predict future values (Viek, 2018). As such, if SAT manufacturing processes are cyclical in nature, LSTM could be effective since it can take previous manufacturing data in consideration to forecast a better estimation for the subsequent production cycle.

On the flipside, for ANN to function effectively, it requires a complete dataset with many data points (Bajic, Cosic, Lazarevic, Sremcev, & Rikalovic, 2018). At present, it is understood that the data collection process at the manufacturing plants are manually inputted which results in many data inconsistencies. On top of that, the data is segmented into many different files and there is no fixed structure. Hence, this complicates the data analysis process and SAT might want to automate this in future.

**Data Collection & Communication**

To improve the data collection process, SAT might consider the usage of sensors and Radio Frequency Identification (RFID) tags to track assembly lines. These tags are able to transit data regarding production quantity, quality of machinery to a central computer network (Boulton, 2014). An example of this successful implementation would be Black & Decker smart factory in Mexico (Journal of Engineering, 2017). Similarly, Black & Decker uses RFID together with WIFI networks to transmit data to a cloud system. This enhances the data collection process and improves productivity. For instance, instead of manually checking for defects this process can be automated with the use of sensors and software. Additionally, this also allows for real-time visibility for manufacturing operations. The data collected can be visualised in real-time into dashboards for the manager to check on the productivity of the factory. On the other hand, such implementations can be costly. In additional to financial costs, implementation challenges also include educating workers/managers about the technology at hand (Boulton, 2014). In spite of this, SAT should explore this option as RFID technology can be installed directly to SAT’s current assembly lines.

References for recommendation

Bajic, B., Cosic, I., Lazarevic, M., Sremcev, N., & Rikalovic, A. (2018). *Machine Learning Techniques for Smart Manufacturing: Applications and Challenges in Industry 4.0.* Serbia: 9th International Scientific and Expert Conference TEAM 2018.

Setyawati, B., Sahirman, S., & Creese, R. (2002). Neural networks for cost estimation. *AACE International Transactions,*ES131-ES139.

Viek, T. (2018, October 24). *LSTM Neural Network for Industry 4.0 Condition-based Monitoring*. Retrieved March 20, 2019, from Towards Data Science: https://towardsdatascience.com/lstm-neural-network-for-industry-4-0-condition-based-monitoring-afee73c7752c

Boulton, C. (2014, August 1). Stanley Black & Decker Retools Factory for the Internet of Things. *Dow Jones Institutional News*.

Stanley Black &Decker; Stanley Black & Decker Opens Stanley Security Futures Innovation Factory In Boston. (2017). *Journal of Engineering*, p. 1044.

References

Visualisation comparison Table X

(Bobriakov, 2018)

[**https://medium.com/activewizards-machine-learning-company/a-comparative-analysis-of-top-6-bi-and-data-visualization-tools-in-2018-658490665973**](https://medium.com/activewizards-machine-learning-company/a-comparative-analysis-of-top-6-bi-and-data-visualization-tools-in-2018-658490665973)

**(Pardo-Bunte, 2018)**

[**https://www.betterbuys.com/bi/reviews/microstrategy-business-intelligence/**](https://www.betterbuys.com/bi/reviews/microstrategy-business-intelligence/)

[**https://www.microstrategy.com/us/product/why-microstrategy**](https://www.microstrategy.com/us/product/why-microstrategy)

**(Mazenko, 2016)**

[**https://www.betterbuys.com/bi/qlikview-pricing/**](https://www.betterbuys.com/bi/qlikview-pricing/)

**About cloud usage**

Tableau support with cloud

[**https://www.tableau.com/about/blog/2016/5/tableaus-approach-cloud-its-all-about-choice-53962**](https://www.tableau.com/about/blog/2016/5/tableaus-approach-cloud-its-all-about-choice-53962)

PowerBI & cloud

[**https://docs.microsoft.com/en-us/power-bi/service-get-data**](https://docs.microsoft.com/en-us/power-bi/service-get-data)

Qlik Sense & cloud

[**https://help.qlik.com/en-US/sense-cloud/Subsystems/CloudHub/Content/Sense\_Hub/LoadData/connect-data-sources.htm**](https://help.qlik.com/en-US/sense-cloud/Subsystems/CloudHub/Content/Sense_Hub/LoadData/connect-data-sources.htm)

Microstrategy w cloud

<https://www.microstrategy.com/us/product/cloud>

Power BI Price: [**https://docs.microsoft.com/en-us/power-bi/service-admin-purchasing-power-bi-pro**](https://docs.microsoft.com/en-us/power-bi/service-admin-purchasing-power-bi-pro)

Qlik Sense pricing

<https://www.qlik.com/us/pricing.>

Tableau price

<https://buy.tableau.com/>

Microstrategy data source support

<https://community.microstrategy.com/s/article/KB327877-Which-databases-data-sources-are-supported-and?language=en_US>

**Marr, 2018**

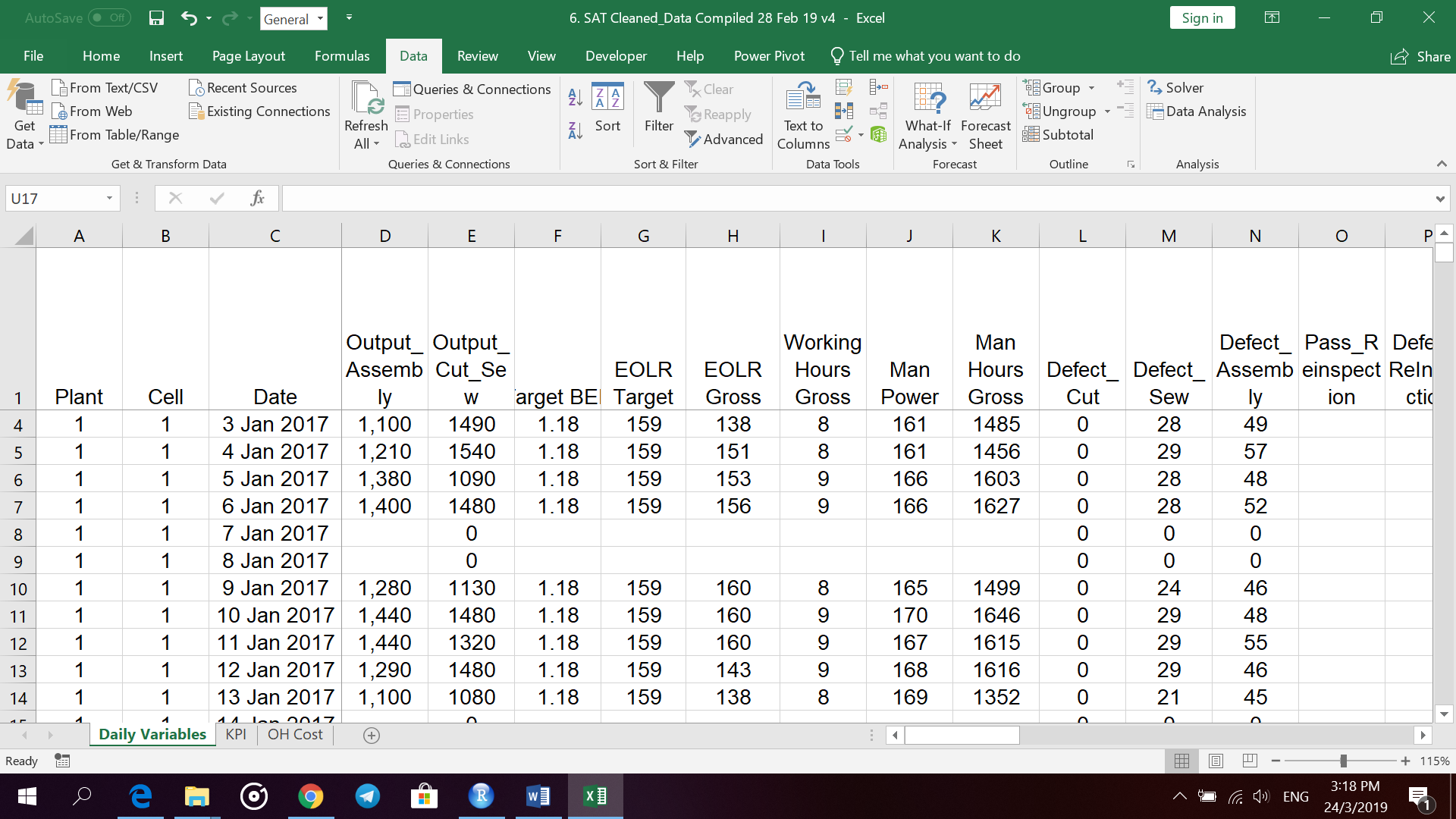
[**https://www.forbes.com/sites/bernardmarr/2018/06/20/comparing-data-visualization-software-here-are-the-7-best-tools-for-2018/#623a0ab26d0b**](https://www.forbes.com/sites/bernardmarr/2018/06/20/comparing-data-visualization-software-here-are-the-7-best-tools-for-2018/#623a0ab26d0b)

**Appendix**

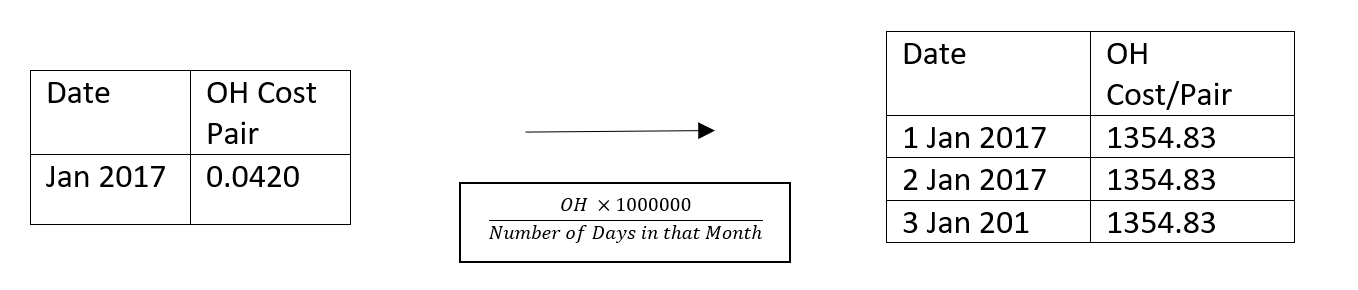
**Appendix A - Data Dictionary**

|  |  |
| --- | --- |
| Data Item | Meaning |
| OutPut Assembly | Total amount of goods fully assembled at the end of the day |
| Target BEP | Target set by management for the Break Even Point |
| EOLR Target | No. of pairs of shoes the line should be producing to be profitable (i.e meet planned CBD) |
| EOLR Gross | Actual pairs of shoes produced per Cell |
| Working Hours Gross | Actual working hours |
| Man Power | Total direct + indirect labour |
| ManHours Gross | Total actual manhours put in production |
| PPH Actual Gross | Output / manhours gross |
| Qty DownTime | Number of times Machine breakdown |
| DownTime Seconds | Duration in seconds of machine breakdown |
| Operating Time Sec | Duration in seconds that machine is running |
| Downtime Percentage | Ratio of Downtime seconds : Operating time sec in percentage |
| Output Cut& Sew | Total Amount of goods that has been cut & sew but have not been assembled |
| Defect Cut | Defects incurred during the cutting process |
| Defect Sew | defects incurred during the sewing process |
| Defect Assembly | defects incurred during the assembly process |
| RFT Cut | Right first time or 1st pass yield during the cutting process |
| RFT Sew | Right first time or 1st pass yield during the sewing process |
| RFT Assembly | Right first time or 1st pass yield during the assembly process |
| RFT Reinspection | Right first time or 1st pass yield during the reinspection process where the defects corrected |

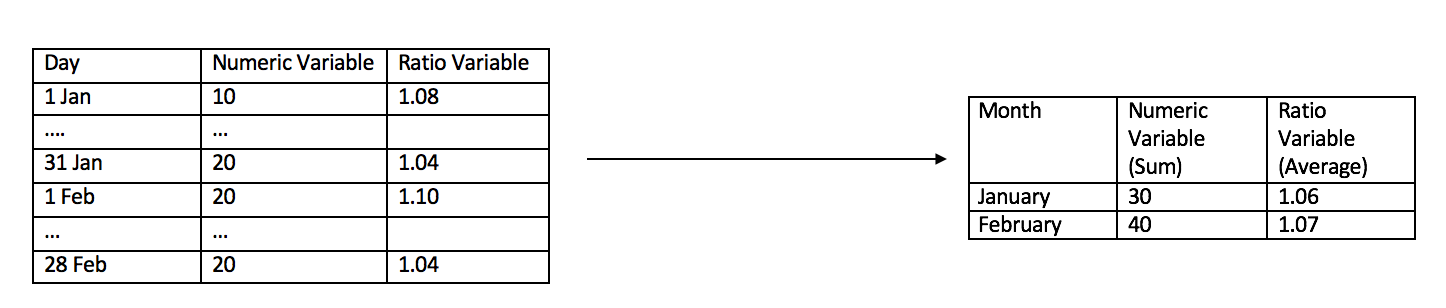
**Appendix B - Snapshot of Data**

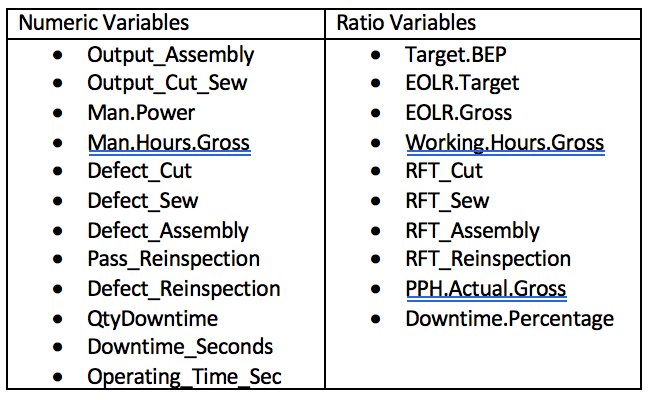
****

**Appendix C - Example of Data Preparation for Daily Approach**

****

**Appendix D - Example of Data Preparation for Monthly Approach**

****

****

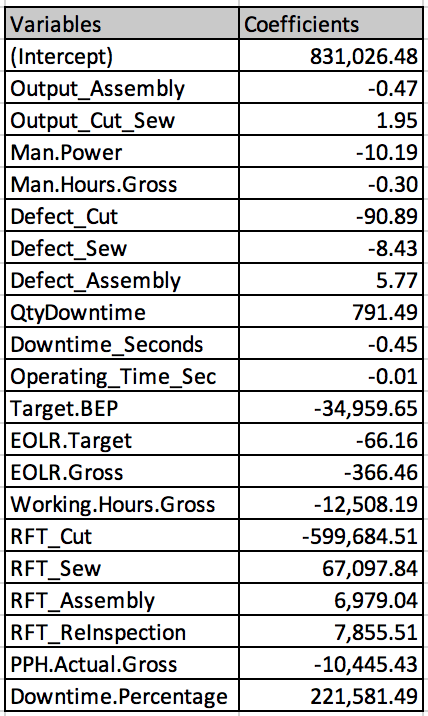
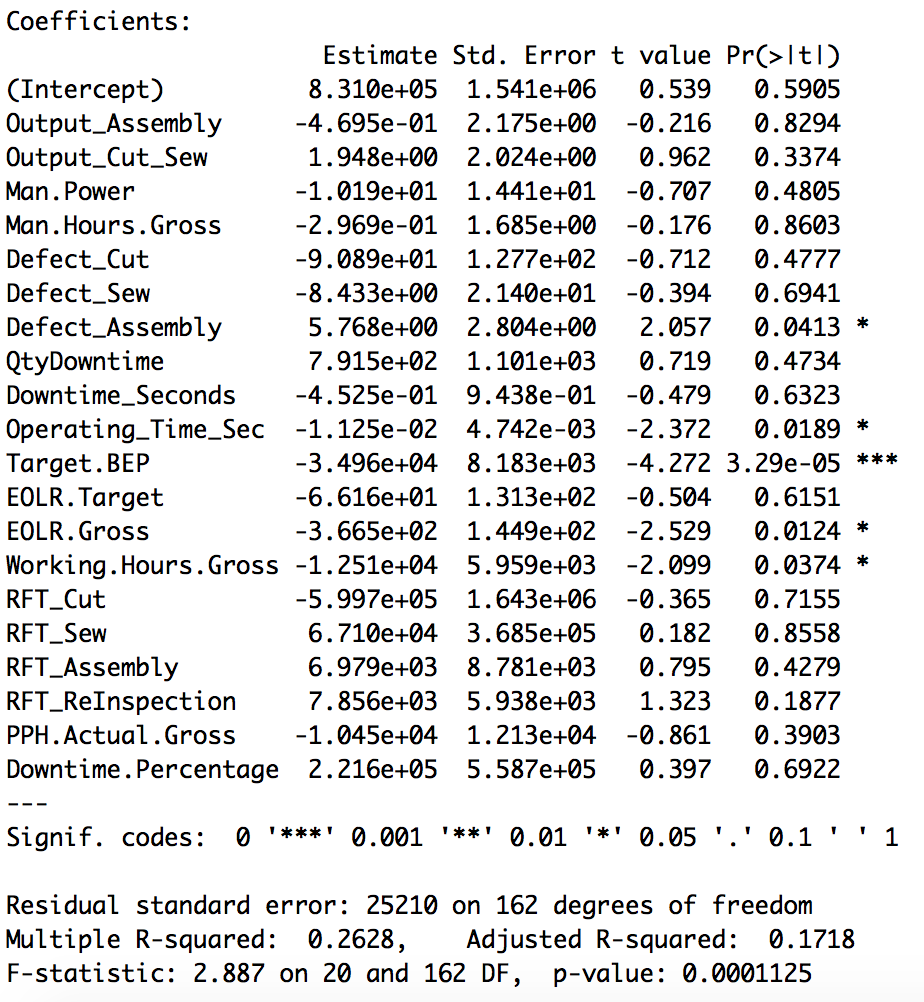
**Appendix X - Results of Linear & LASSO Model**

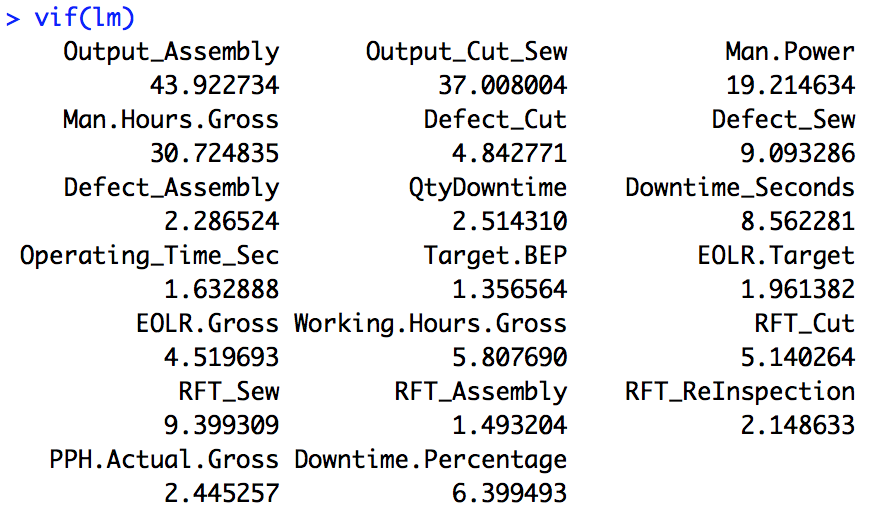
**Appendix X - Model Results Monthly**

**Predicting Total Overhead Costs**

**Linear Regression Model 1:**

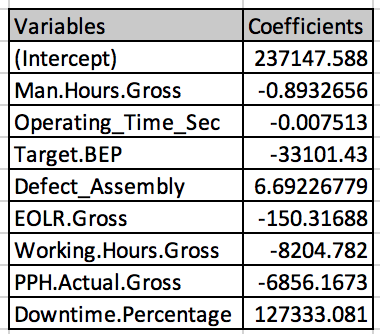
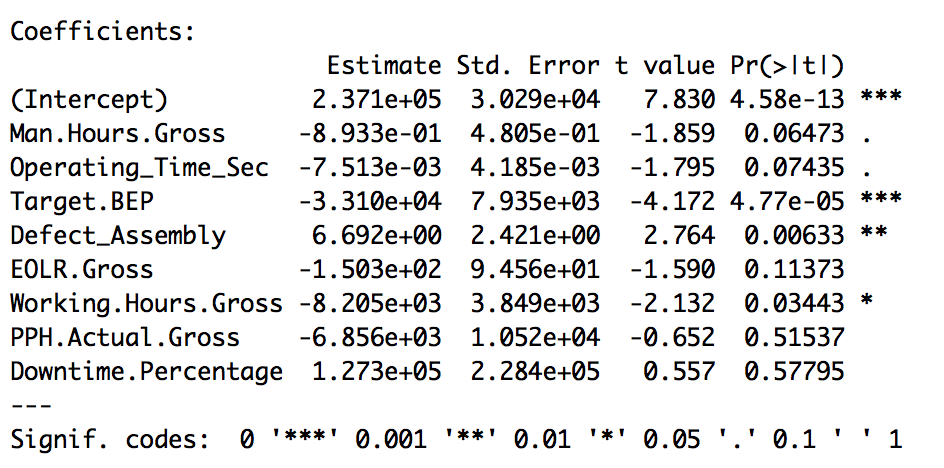
Includes all variables

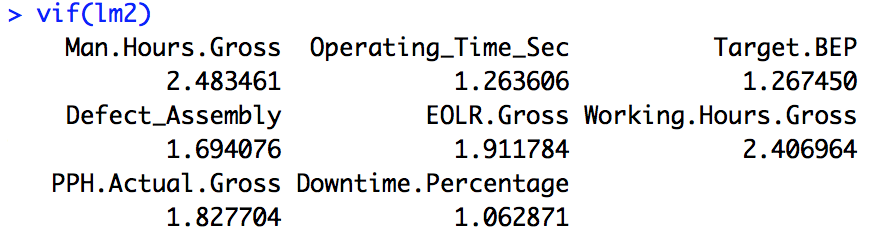


Variable Inflation Factor:

Out of sample MAE: 103590.1

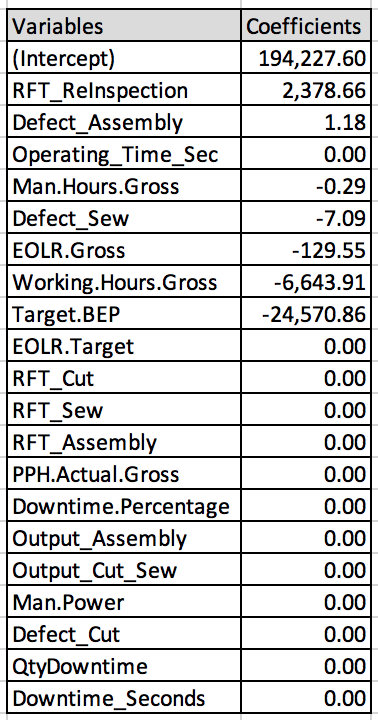
**Final Linear Regression**

****

Variable Inflation Factor:

MAE: 15804.55

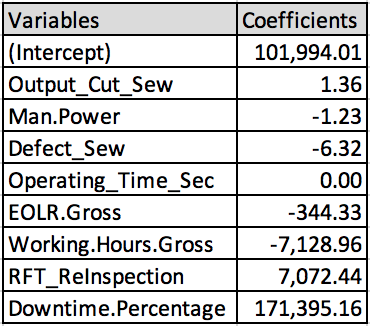
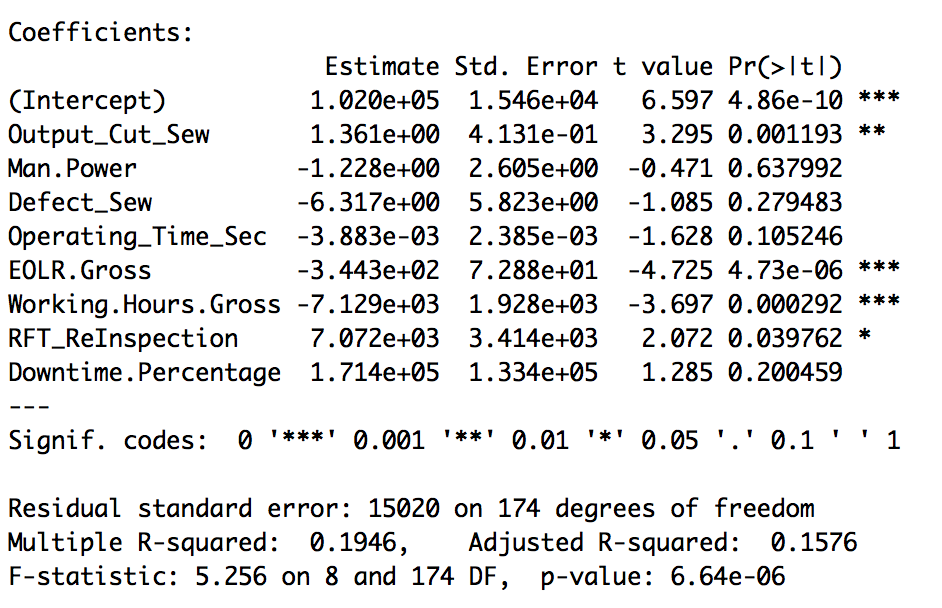
**LASSO model**

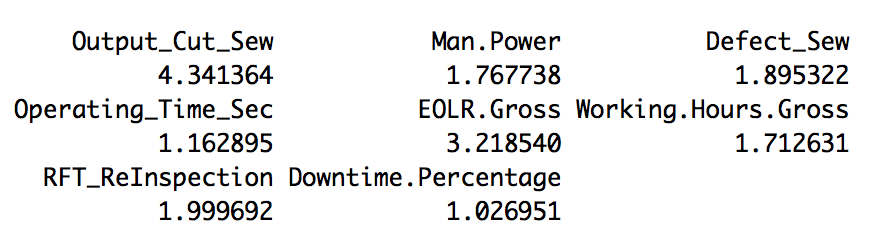


MAE: 26915.22

**Predicting variable overhead costs**

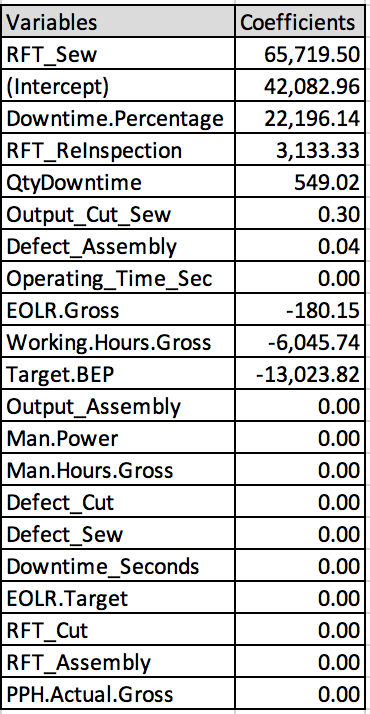
**Final Linear Regression**

****

Variable Inflation Factor: 

MAE: 4936.385

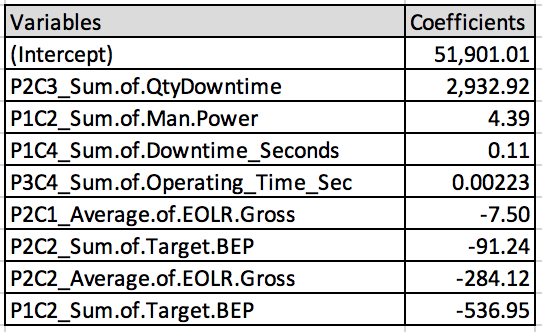
**LASSO**



MAE: 3742.755

**Predicting variable overhead costs using non-aggregated data**

**LASSO**

****

MAE: 10451.69